THE UNIVERSITY *of* EDINBURGH

Estimating the host genetic contribution to the epidemiology of infectious diseases

Debby Lipschutz-Powell

PhD Quantitative Genetics – The University of Edinburgh – 2014

Contents

Figures and Tables

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Date: 11/06/2014

Signed: 

Debby Lipschutz-Powell

Acknowledgements

First of all, I would like to thank my supervisors Andrea Doeschl-Wilson and John Woolliams for all their expertise, help and support throughout my PhD. I would also like to thank Piter Bijma for his insights and supervision with regards to my work on Indirect Genetic Effects and Luís-Alberto García-Cortés for his original derivation of the posterior distribution used in Chapter 5 and all his insights and supervision with regards to the work developed in Chapter 5. This thesis would not have been possible without them.

I would like to thank all the members of the "Bayesian inference for epidemiology" group for their valuable insights. In particular, Osvaldo Anacleto has really helped to improve my knowledge with regards to Bayesian Inference and I would like to thank him for his time and patience with me in the last year. A special thanks also goes to Ricardo Pong-Wong who taught me how to program, is always there to help and give advice and with whom I've had some very interesting discussions.

This PhD was funded by the BBSRC and Cobb-Vantress Incorporated within the remit of a Bioscience KTN Industrial CASE studentship. In particular I would like to thank Gosse Veninga, Rachel Hawken, Randy Borg and Mitchell Abrahamsen at Cobb Vantress for all their interest and insights with regards to this project.

I would like to thank Valentina Riggio and Zeenath Islam for sacrificing their week-end to proof-read this thesis.

Last but not least, I would like to thank my husband, parents and siblings and the "cake list" for their everyday friendly support throughout my PhD.

Abstract

Reducing disease prevalence through selection for host resistance offers a desirable alternative to chemical treatment which is a potential environmental concern due to run-off, and sometimes only offers limited protection due to pathogen resistance for example (Chen et al., 2010). Genetic analyses require large sample sizes and hence disease phenotypes often need to be obtained from field data. Disease data from field studies is often binary, indicating whether an individual became infected or not following exposure to infectious pathogens. In genetic analyses of binary disease data, however, exposure is often considered as an environmental constant and thus potential variation in host infectivity is ignored. Host infectivity is the propensity of an infected individual to infect others. The lack of attention to genetic variation in infectivity stands in contrast to its important role in epidemiology.

The theory of indirect genetic effects (IGE), also known as associative or social genetic effects, provides a promising framework to account for genetic variation in infectivity as it investigates heritable effects of an individual on the trait value of another individual. Chapter 2 examines to what extent genetic variance in infectivity/susceptibility is captured by a conventional model versus an IGE model. The results show that, unlike a conventional model, which does not capture the variation in infectivity when it is present in the data, a model which takes IGEs into account captures some, though not all, of the inherent genetic variation in infectivity. The results also show that genetic evaluations that incorporate variation in infectivity can increase response to selection and reduce future disease risk. However, the results of this study also reveal severe shortcomings in using the standard IGE model to estimate genetic variance in infectivity caused by ignoring dynamic aspects of disease transmission.

Chapter 3 explores to what extent the standard IGE model could be adapted for use with binary infectious disease data taking account of dynamic properties within the remit of a conventional quantitative genetics mixed model framework and software. The effect of including disease dynamics in this way was assessed by comparing the

accuracy, bias and impact for estimates obtained for simulated binary disease data with two such adjusted IGE models, with the Standard IGE model. In the first adjusted model, the Case model, it was assumed that only infected individuals have an indirect effect on their group mates. In the second adjusted IGE model, the Case-ordered model, it was assumed that only infected individuals exert an indirect effect on susceptible group mates only. The results show that taking the disease status of individuals into account, by using the Case model, considerably improves the bias, accuracy and impact of genetic infectivity estimates from binary disease data compared to the Standard IGE model. However, although heuristically one would assume that the Case-ordered model would provide the best estimates, as it takes the disease dynamics into account, in fact it provides the worst. Moreover, the results suggest that further improvements would be necessary in order to achieve sufficiently reliable infectivity estimates, and point to inadequacy of the statistical model.

In order to derive an appropriate relationship between the observed binary disease trait and underlying susceptibility and infectivity, epidemiological theory was combined with quantitative genetics theory to expand the existing framework in Chapter 4. This involved the derivation of a genetic-epidemiological function which takes dynamic expression of susceptibility and infectivity into account. When used to predict the outcome of simulated data it proved to be a good fit for the probability of an individual to become infected given its own susceptibility and the infectivity of its group mates. Using the derived function it was demonstrated that the use of a linear IGE model would result in biased estimates of susceptibility and infectivity as observed in Chapters 2 & 3.

Following the results of Chapter 4, the derived expression was used to develop a Markov Chain Monte Carlo (MCMC) algorithm in order to estimate breeding values in susceptibility and infectivity in Chapter 5. The MCMC algorithm was evaluated with simulated disease data. Prior to implementing this algorithm with real disease data an adequate experimental design must be determined. The results suggest that there is a trade-off for the ability to estimate susceptibility and infectivity with

regards to group size; this is in line with findings for IGE models. A possible compromise would be to place relatives in both larger and smaller groups. The general discussion addresses such questions regarding experimental design and possible areas for improvement of the algorithm.

In conclusion, the thesis advances and develops a novel approach to the analysis of binary infectious disease data, which makes it possible to capture genetic variation in both host susceptibility and infectivity. This approach has been refined to make those estimates increasingly accurate. These breeding values will provide novel opportunities for genome wide association studies and may lead to novel genetic disease control strategies tackling not only host resistance but also the ability to transmit infectious agents.

Chapter 1. Introduction and literature review

1.1 Control of infectious disease in livestock: an epidemiological perspective

As was seen in the foot and mouth epidemic, infectious diseases in livestock can have a significant impact on the sustainability of livestock production. Moreover, the need to contain epidemics has been further emphasized by the threat of transmission to other species as illustrated in the recent swine-origin H1N1 influenza (Dawood *et al.* 2009) and avian influenza H5N1 (Salzberg *et al.* 2007) epidemics. Epidemiology is the study of the transmission and control of epidemic disease (1992). The use of mathematical models to assess the effectiveness of control strategies has been common in epidemiology since the 18th century when Bernoulli published a model to assess the risks and benefits of inoculation against smallpox (Anderson and May 2006). Typically epidemiological models in modern times assume homogeneous populations and describe movement from one disease category to another, following average rates, through the use of differential equations. Although these models differ widely in scope and approach they all tend to focus on the reduction of the average amount of transmission between infectious and susceptible individuals as measured by the basic reproduction number R_0 (Keeling and Rohani 2008). R_0 is the expected number of secondary infections an infected individual may cause in a fully susceptible population (Keeling and Rohani 2008).

The basic reproduction number R_0 may be reduced either by reducing the exposure of susceptible individuals to the pathogens and/or reducing the susceptibility of these individuals. Susceptibility is the probability of susceptible individuals to become infected upon exposure to a unit dose of pathogens. Susceptibility may be reduced either through interventions such as vaccination or selective breeding. However, vaccines are not always available for biological and/or financial reasons and sometimes cause selection for increased pathogen virulence (Gimeno 2008; Kimman *et al.* 2009). Moreover, DIVA (Differentiating Infected and Vaccinated Animals) vaccines are not always available and vaccines sometimes offer only partial

protection rendering it difficult to detect infected individuals (Van Oirschot 2001). The merits of breeding as part of infectious disease control strategies are discussed in section 1.2.

Reducing the exposure may be achieved, for example, by reducing the virulence or number of pathogens through chemical treatment, such as antibiotics. However, chemical treatments are not always available or desirable due to financial reasons, potential environmental concern due to run-off, and potential for pathogen resistance (Chen *et al.* 2010; Demeler *et al.* 2010). Another method to reduce exposure, is to limit the propensity of infected individuals to transmit the infection, i.e. their infectivity. This may be achieved for example through isolation of infected individuals and/or individuals who came into contact with them and is inherent to many control strategies. For example, legislation imposing a standstill for a number of days between moving cattle on and off a farm has been shown to drastically impact the potential spread of Foot and Mouth Disease (Vernon and Keeling 2012). However, the success of such strategies is dependent on a reliable and early detection of infected individuals which is not always feasible (Charleston *et al.* 2011). Moreover, culling of (potentially) infected individuals is often part of such control strategies and is very costly. Furthermore, these strategies may be compromised by ongoing contact with a wildlife reservoir as is thought to be the case for Bovine Tuberculosis in the UK due to the presence of infected badgers (Krebs *et al.* 1998). Culling of the wildlife reservoir is expensive, not always effective and may clash with conservation policies and public opinion (Jenkins *et al.* 2010).

A combination, targeting both exposure and susceptibility, which has increasingly been explored is to target individuals with an inherently high infectivity for vaccination and/or treatment. Indeed, the distribution of infectivity has been shown to often be highly skewed with a small proportion of individuals accounting for a large proportion of transmission events (Lloyd-Smith *et al.* 2005). Such super-spreading individuals can have a profound effect on the risk and severity of an outbreak, as was shown in the recent SARS epidemic (Shen *et al.* 2004).

1.2 Reduction of infectious disease prevalence through selection

Given that epidemiological studies advocate control strategies which focus on the reduction of both susceptibility and infectivity, it would be interesting to quantify whether this could be achieved through genetic selection. However, currently efforts to control infectious diseases through selective breeding tend to focus on disease resistance. Nonetheless, some of the traits identified as indicators of resistance are dependent on both the individual's susceptibility and infectivity such as Faecal Egg Counts (FEC) for nematode infections in sheep. Indeed, different host immune responses in adult sheep were shown to reduce *O. circumcicta* FEC through reduction in worm burden and worm fecundity respectively (Stear *et al.* 1997). The fact that selecting for reduced FEC affects both susceptibility and infectivity was noted by Bishop and Stear (1997) as they first demonstrated that taking the epidemiology into account one might predict an actual response to selection which is much greater than that predicted using quantitative genetic theory alone. Nonetheless, there seems to be little information regarding whether infectivity is genetically controlled. Moreover disease traits such as FEC for nematode infections or somatic cell counts for mastitis are only proxies for susceptibility and infectivity. Thus appropriate analytical tools are required to infer susceptibility and infectivity from disease traits.

Selective breeding for disease resistance based on estimated breeding values from pedigreed data has been part of control strategies for a range of diseases such as mastitis in dairy cattle (Heringstad *et al.* 2000) and nematode infections in sheep (Bishop and Morris 2007). However, large phenotyped datasets across generations are required for genetic selection and thus data tends to originate from the field rather than challenge studies for both economical and ethical reasons. There are several issues with using field data, however. It is rather noisy due to imperfect sensitivity and specificity of diagnostic tests and lack of knowledge regarding time of infection

and exposure of individuals. This noise often results in poor heritability estimates (Bishop and Woolliams 2010). Furthermore, genetic interpretation becomes more difficult with each generation as reducing disease prevalence through breeding for disease resistance results in a reduction in the number of exposed individuals. In this way, the development of high-throughput genomic tools has made genetic selection an increasingly desirable and attainable complement to conventional disease control strategies. Indeed, Genome Wide Selection, i.e. selection on the sum of the effect of all Single Nucleotide Polymorphisms (SNPs) on the trait of interest, is now feasible as dense SNP arrays in excess of 50k are available for most farm animal species. It is particularly attractive as pedigree knowledge and extensive phenotyping and genotyping of each generation is not required. In other words, for disease traits, extensive challenge studies of each generation are not required. Nevertheless, many thousands of animals are still required in order to identify the SNPs associated with the trait of interest in the first instance. An important step in bringing the lessons from epidemiology across to the field of genetics would therefore be to develop an analytical tool to estimate breeding values in both susceptibility and infectivity from field host disease data. As previously mentioned, for reasons of feasibility field data regarding infectious diseases often comes in binary form, merely indicating whether an individual became infected or not by/at the time of measurement. Quantitative genetic analysis of binary data however is not without its issues, as outlined in the next section and revealed throughout this thesis.

1.2.1 Current quantitative genetic analysis of binary data

In order to estimate breeding values from binary data, it is usually assumed that the observed binary outcome is the result of an underlying continuous trait, called the liability, which is linear for genetic and environmental effects, and an error term. Breeding values may then be estimated for this linear underlying trait through the use of a mixed model which allows simultaneous estimation of fixed effects and random genetic parameters (Falconer and Mackay 1996). Generalized Linear Mixed Models (GLMM) scale estimates obtained with such a Linear Mixed Model to the observed

scale using a (non-linear) link function. There are publically available tools which use either a frequentist approach, e.g. ASReml (Gilmour *et al.* 2006), or a Bayesian approach, e.g. (Tsuruta and Misztal 2006), to estimate genetic parameters using such GLMMs.

The link functions more commonly used for binary traits are the probit, logit and complementary log-log link functions (Khuri *et al.* 2006). It is assumed that the liability follows a cumulative logistic and a cumulative normal distribution for the logit and probit link functions respectively. These distributions are both smooth symmetrical sigmoids and are sometimes used as approximations of each other. Both distributions are often used due to ease of implementation into computational codes. Moreover, the logit link function is also used for ease of interpretation because if the liability represents the probability of success, then the logit function is the logarithm of the odds ratio for success. The probit link function is often used due to the attractive properties of the normal distribution. Moreover, it is often used in conjunction with the assumption that all individuals which exceed a threshold liability value will have a binary observation equal to one, i.e. a threshold model. It is customary to assume equal exposure when threshold models are used to analyse disease data. Thus differences in disease status due to the underlying liability are equated to differences in susceptibility only. This assumption could be lifted under the threshold model but this would require the assumption that highly resistant individuals given sufficient exposure will become infected. Finding a relationship which fulfils this requirement as well as being biologically relevant and where differences of exposure may be disentangled from susceptibility, may not be straightforward.

Another commonly used link function to analyse binary data is the complementary log-log link function (Khuri *et al.* 2006). This function is asymmetrical with a sharper rate of increase as it approaches one. It is notable for its parallels with survival analysis as the liability is equivalent to the natural logarithm of the cumulative hazard function. In terms of infectious disease this is tantamount to saying that the liability at a given time t is the natural logarithm of the per capita rate

of infection i.e. the force of infection. However, individual variation in the rate at which susceptible individuals become infected is traditionally assumed to be due to susceptibility only in Quantitative Genetic analysis.

If one has data such as time to infection from a challenge study for example, tools have been developed, such as 'Survival Kit' (Ducrocq *et al.* 2010) to estimate genetic parameters for time to infection or death through infection through survival analysis. This approach is particularly useful if it is not possible to eradicate the disease and interest lies in breeding for tolerance instead. In that situation all individuals would be challenged equally at the start of the experiment and time to death would be recorded. For transmission dynamics, however, the interesting phenotype would be time to infection. However, this may not always be feasible to obtain as the onset of infection is usually unknown. In field conditions repeated binary data may possibly be used as a reliable estimate for time to infection in a similar manner to survival scores. Survival scores are sequential binary data, indicating whether an individual has survived or not to given time points. Recently Ødegård et al. (2011) developed a type of survival model, called a cure model, for use with a Gibbs sampler to estimate genetic parameters from survival scores. Cure models are survival models which take into account the fact that a fraction of individuals may not be susceptible and would therefore never become infected. These models may be utilised to separate individuals which neither die nor show signs of disease due to resistance rather than tolerance. However, this model also assumes equal or random exposure. The assumption of equal or random exposure is likely not to be met in the case of infectious disease, however. Indeed, in the case of infectious diseases exposure is directly dependent on the number of infectious individuals or the infective compunds an individual may come into contact with. This number will vary as the number of infectious individuals changes over time and is likely to differ from one individual to another. Moreover, should there be variation in infectivity then each contact would not carry the same weight in terms of exposure.

Furthermore, it is worth noting that the use of GLMMs with binary data is not without its technical complications. For example, ASReml (Gilmour *et al.* 2006)

uses a method called Penalized Quasi Likelihood in order to estimate the genetic parameters. This method uses a first order Taylor series approximation to the likelihood and may therefore be prone to bias as has been widely documented (e.g. Breslow and Lin 1995; Rodriguez and Goldman 2001). Alternative approaches to analysis and inference use Bayesian methods, and these have been found to suffer from two major problems when analysing binary data. The Extreme Category Problem as described among others by Misztal et al. (1989), which occurs when all individuals with a given fixed effect have the same binary outcome, and biased estimates when using an animal model (e.g. Luo *et al.* 2001). Algorithms, for threshold models only, have recently been developed to overcome these problems in most situations (e.g. Ødegård *et al.* 2010a). Extensive literature searching has only identified tools for estimating genetic parameters from binary data with a link function that do not suffer from bias in the case of threshold models. However, as mentioned earlier, threshold models are not necessarily adequate for capturing genetic variation in disease resistance in natural disease outbreaks due to variation in exposure and epidemiological interpretation. Thus, given the difficulties involved in using GLMMs to estimate genetic parameters from binary data, it is often assumed that the liability at the observed level follows a linear model. It is worth noting that when a threshold model is adequate, breeding values estimated with both a linear and a threshold model have been shown to agree well (Heringstad *et al.* 2003; Ødegård *et al.* 2010b). Therefore, in the absence of more adequate tools, using a linear model on the observed level, and potentially transforming the estimates after the analysis, may well be the most robust method and is therefore employed in Chapters 2 and 3. Chapter 4 then explores analytically which relationship would be adequate in order to obtain estimates for susceptibility and/or infectivity from binary disease data.

1.2.2 Capturing genetic variation in susceptibility and infectivity using IGEs

There are two major challenges in estimating breeding values in susceptibility and infectivity from disease data collected from field studies. Firstly, infectivity is

difficult to measure directly as its effect is observed in a different individual from the one expressing it. Secondly, expression of susceptibility and infectivity depends on infection status, i.e. only susceptible individuals express susceptibility and infectious individuals express infectivity. In other words, expression of susceptibility and infectivity is dynamic.

The problem of infectivity being observed in a different individual than the one expressing it may be resolved through the use of an Indirect Genetic Effects model. Indeed, over the last forty years, the theory of Indirect Genetic Effects (IGE) has been developed to investigate the impact of interactions among individuals on the expression and evolution of traits (Bijma *et al.* 2007a; Griffing 1967; Moore *et al.* 1997; Muir 2005). This thesis includes the first studies to examine the application of IGE models to study host genetic influence underlying infectious disease.

1.2.2.1 Indirect Genetic Effects

Classically, the phenotypic trait value of an individual i (P_i) is partitioned into an additive genetic value (A_i), or breeding value, and an environmental deviation (E_i) (equation (1.1)) (Falconer and Mackay 1996). The environmental deviation consists of all factors affecting the trait which are external to the individual, such as food or temperature, and non-additive genetic components i.e. dominance and epistasis. Such external factors could be due to interactions with other individuals. In this way, the environmental deviation can be further partitioned into a 'direct' environmental deviation ($E_{D,i}$) which is not caused by other individuals, and the sum of the effects of each group member j ($P_{S,j}$) (equation (1.2)) (e.g. Bijma *et al.* 2007a; Griffing 1967). This last part of the environmental deviation has been termed 'associative effects', 'social effects', 'indirect genetic effects' (IGE) or 'heritable environment' as it is external to the focal individual but may have a genetic component (equation (1.3)) e.g. (Bijma *et al.* 2007a). Thus an IGE is a heritable effect of an individual on the trait value of another.

$$P_i = A_i + E_i \tag{1.1}$$

$$P_i = A_{D,i} + E_{D,i} + \sum_{j=1}^{n} P_{S,j} \qquad j \neq i \tag{1.2}$$

$$= A_{D,i} + E_{D,i} + \sum_{j=1}^{n} (A_{S,j} + E_{S,j}) \qquad j \neq i \tag{1.3}$$

Two different approaches were developed to estimate response to selection taking IGEs into account. The first statistical approach was developed by Griffing in a series of papers from 1967 to 1982 (Griffing 1967; 1968a; 1968b; 1969; 1976a; 1976b; 1981a; 1981b; 1981c; 1981d; 1982a; 1982b; 1982c; 1982d) aiming to investigate competitive interactions among crop plants. In his models Griffing defines the total breeding value of an individual (A_T) as the sum of the direct additive genetic value ($A_{D,i}$), i.e. part of the phenotypic value determined by the individual's own genes, and the total effect of genetic origin an individual has on the expression of the selected trait in its group members ($A_{S,i}$) (equation (1.4)).

$$A_T = A_{D,i} + (n-1)A_{S,i} \tag{1.4}$$

He then estimated the change in mean phenotype, or response to selection, by deriving the relevant expression for relative fitness depending on group composition and selection criterion. In this way, Griffing's models treat a different specific scenario in each article rendering comparisons of different experimental designs difficult. More recently, Bijma et al. (Bijma *et al.* 2007a; 2007b; Bijma and Wade 2008) continued Griffing's work and developed a generic expression using the Price equation (equation (1.5)). Indeed, using equation (1.5) response to selection ($\Delta\bar{P}$) taking IGEs into account can be estimated for any level of selection (g), e.g. mass selection g=0, group selection g=1, on groups of any size (n) and any degree of relatedness (r).

$$\Delta\bar{P} = \beta[(g + r + (n-2)gr)Var(A_T) + (1-g)(1-r)(Var(A_D) + (n-1)Cov(A_D, A_S))] \tag{1.5}$$

Where β represents the direct selection gradient or regression coefficient of the phenotypic value of an individual on its breeding value.

The second approach, stemming from maternal effect models by Falconer (Falconer 1965) and Kirkpatrick & Lande (Kirkpatrick and Lande 1989; Lande and Kirkpatrick 1990), was developed by Moore et al. in 1997 (Moore *et al.* 1997; Wolf *et al.* 1998; Wolf *et al.* 1999). The main two differences between this more functional approach and the statistical approach pioneered by Griffing are that the trait generating the IGE (P') is a known and measurable trait and the interaction occurs between two specific individuals. The extent to which that trait affects the expression of the selected trait is then specified by a regression coefficient (ψ_{ab}) (system of equations (1.6) & (1.7)).

$$P_i^a = A^a + E^a + \psi_{ab} P_j^b \quad (1.6)$$

$$P_j^b = A^b + E^b + \psi_{ba} P_i^a \quad (1.7)$$

With a and b referring to traits and i and j to individuals. Note that this expression is equivalent to that in equation (1.2) if the IGE is generated by only one trait P^b in a group of two individuals. Moore et al.'s model however includes the possibility of the indirect trait P^b to be affected in return by the trait P^a (equation (1.7)). The response to selection is then obtained the following way.

$$\Delta \bar{P}^a = \frac{1}{(1-\psi_{ab}\psi_{ba})^2}[(Var(A^a) + \psi_{ab}Cov(A^a, A^b)\beta_a + (Cov(A^a, A^b) + \psi_{ba}Var(A^b))\beta_b] \quad (1.8)$$

With β_a and β_b representing the selection gradients on traits P^a and P^b respectively. Several traits can be considered simultaneously by replacing the coefficients by vectors and the variances and covariances by their respective matrices.

The purpose of Moore et al.'s functional approach is to examine the effect that one trait has on the evolution of another. Griffing's statistical approach, on the other

hand, aims to estimate response to selection for breeding purposes. To this effect the IGE is a variance component to be estimated using statistical techniques such as REML; no assumptions are made as to the nature of the trait and it does not need to be measurable. Given that susceptibility and infectivity may be hard to measure in some cases, the use of the statistical approach may be more appropriate. Should both traits be directly measurable, the functional approach may provide interesting insights in the relationship between infectivity and susceptibility. Also, estimating values for the regression coefficient isn't always straightforward. To this purpose McGlothlin and Brodie (2009) proposed to estimate the genetic (co)variance components using the statistical approach and then estimate the regression coefficients (ψ_{ij}) using the identity, $\mathbf{\Psi} = \mathbf{G}_{SD}\mathbf{G}_{D}^{-}$ with $\mathbf{\Psi}$ being a matrix of regression coefficients, $\mathbf{G}_{SD}$ the indirect genetic-direct effects covariance matrix and $\mathbf{G}_{D}$ the direct effects variance matrix in groups of two individuals. In the same article they extend the functional approach to larger groups of size *n* and state that in that case $\mathbf{\Psi} = \mathbf{G}_{SD}[\mathbf{G}_{D} + (n-2)\mathbf{G}_{SD}]^{-1}$. However, as mentioned by McGlothlin and Brodie (2009) simultaneous analysis of all traits is required in order to avoid missing interactions as they may cancel each other out. Knowledge of all traits affecting disease transmission may not be feasible and therefore the statistical method may be more appropriate.

As previously seen, response to selection in both approaches depends on the covariance between the direct and indirect effect/trait (equations (1.5) & (1.8)). In this way, when the direct effects and indirect effects are positively correlated, response to selection will exceed expectations from the classical model without indirect effects. Such discrepancies have been observed by Bishop and Stear (1997) in a model of nematode infection in sheep, where selection for reduced FEC resulted in an observed response 1.7 times greater than expected. In terms of disease, a positive covariance would occur if more resistant individuals are less infectious. Given that only infected individuals can be infectious this appears somewhat intuitive.

When the direct and indirect genetic effects are sufficiently negatively correlated, on the other hand, response to selection could be going in the opposite direction of selection. In disease terms, an apparent negative covariance would occur if individuals showing less clinical signs are considered as more resistant and turn out to be more infectious e.g. by exhibiting a prolonged asymptomatic carrier phase. Another example of negative covariance would occur if individuals who do not easily become infected have poorer mechanisms for clearing the infection and end up being infected and thus possibly also infectious for a longer period. This may occur due to a trade-off in resource allocation (Rauw *et al.* 1998).

Better knowledge of the role of indirect effects on disease prevalence would allow for better design of breeding programmes. For example, relatedness between group members and/or selection based on group rather than individual performance was shown in theory to reduce the importance of the correlation of effects (equation (1.5)), and therefore reduce the risk of response in the wrong direction (Griffing 1981b; Bijma and Wade 2008). These results were corroborated in a fourteen year selection study on broiler behavioural traits by Muir and associates, reviewed by Muir & Craig (1998) as well as several more recent studies (e.g. Bergsma *et al.* 2008; Ellen *et al.* 2008). Overall, implementation of IGE models is currently subject to much research, but to date this is the first study to apply these models to epidemiological characteristics.

1.2.2.2 Applying IGEs to the epidemiological framework.

In epidemiological models host populations are divided into categories depending on their disease state (e.g. susceptible, infected). Transition from one category to another is usually assumed to occur at a constant rate (Anderson & May 2006). This discussion will focus on the relatively simple Susceptible-Infected-Recovered/Removed (SIR) model. The population is divided into susceptible individuals S which have not been infected yet, infected individuals I and recovered individuals R. Infection occurs according to a transmission rate β and recovery at a

rate γ. The transmission rate is a function of infectivity and susceptibility. Traditionally epidemiological models assume homogeneous populations, but an increasing number of models allow heterogeneity in epidemiological parameters (Doeschl-Wilson *et al.* 2011; Lloyd-Smith *et al.* 2006; Mackenzie and Bishop 2001; Nath *et al.* 2008).

Direct and indirect effects on prevalence may be mapped onto the above described SIR model as follows. The impact of an infectious individual on group prevalence depends on (i) the individual's ability to transmit pathogens, which is an IGE, (ii) the susceptibility of the population's members, which is a direct effect. Hence, the effect of an individual on group prevalence depends on expression of indirect and direct effects, and their relative importance depends on disease status (Figure 1).

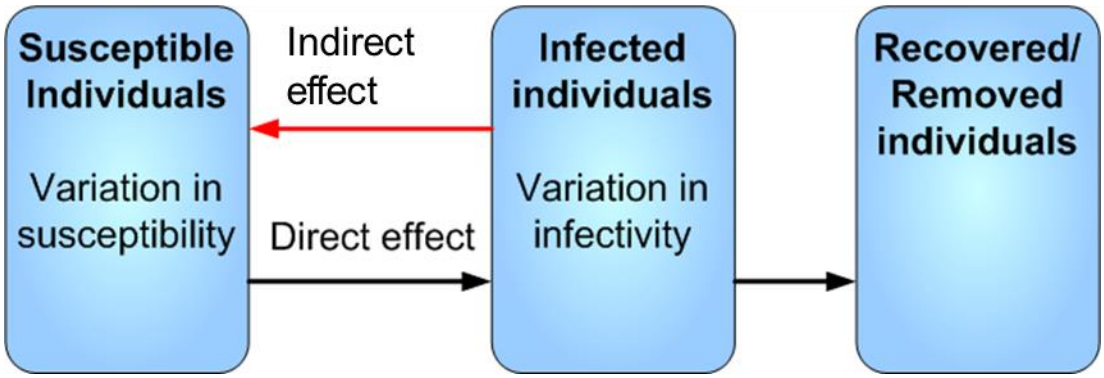

Figure 1-1 Disease dynamics according to an epidemiological SIR model and expression of direct and indirect effects

Thus, expression of indirect and direct effects is expected to change over the time course of infection. To illustrate this, consider a simple SI model i.e. assuming no recovery, where S and I represent the proportion of susceptible and infected individuals respectively. In such a model change in number of infected individuals over time is proportional to both the proportion of infected and susceptible individuals, $dI/dt \propto IS$. As the variation in the indirect effect is expected to be expressed in infected individuals (Figure 1-1), we can expect the variation in IGEs to

be scaled by $S^2 = (1 - I)^2$. Similarly, the variation in direct effects is expected to be expressed in susceptible individuals and will therefore be scaled by I^2 (Figure 1-2). Therefore, as S and I change over the time course of infection, direct and indirect genetic effect variances are expected to change as well. Hence, this illustrates the need to take disease dynamics into account when evaluating response to selection.

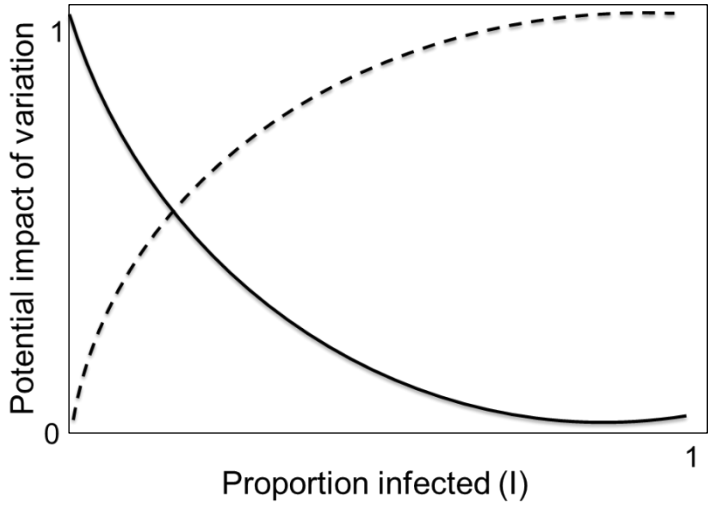

Figure 1-2 Potential impact of individuals' direct and indirect variation depending on prevalence in an SI model
__Indirect, - - Direct

1.3 Project aims and objectives

The aim of this project is to develop a tool which can estimate breeding values in both host susceptibility and infectivity from binary field infectious disease data. To this purpose we investigate in Chapter 2 whether it is possible to capture genetic variation in infectivity from binary disease data with a conventional animal model, and to test whether an IGE model is more appropriate. In Chapter 3 we examine whether it is possible to improve the bias, accuracy and selection response of an IGE model used to estimate breeding values in host susceptibility and infectivity from binary infectious disease data by including disease status and dynamics. In Chapter 4 we derive from first principles an expression linking binary observations to susceptibility and infectivity, incorporating the underlying disease dynamics and

discuss implications for the genetic analysis of susceptibility and infectivity. We then use this expression to develop a Markov Chain Monte Carlo (MCMC) algorithm, detailed in Chapter 5, which estimates breeding values in susceptibility and infectivity from sequential binary data. Finally, means to improve the algorithm developed in Chapter 5 and to implement it to real data as well as future opportunities are discussed in Chapter 6.

1.4 References

1992, pp. in *New Webster's Dictionary and Thesaurus of the English Language*, edited by C. B. S. Lexicon Publications, Inc., Danbury, CT.

Anderson, R. M., and R. M. May, 2006 *Infectious Diseases of Humans*. Oxford University Press, Oxford.

Bergsma, R., E. Kanis, E. F. Knol and P. Bijma, 2008 The contribution of social effects to heritable variation in finishing traits of domestic pigs (Sus scrofa). Genetics **178:** 1559-1570.

Bijma, P., W. A. Muir and J. a. M. Van Arendonk, 2007a Multilevel selection 1: Quantitative genetics of inheritance and response to selection. Genetics **175:** 277-288.

Bijma, P., W. M. Muir, E. D. Ellen, J. B. Wolf and J. a. M. Van Arendonk, 2007b Multilevel selection 2: Estimating the genetic parameters determining inheritance and response to selection. Genetics **175:** 289-299.

Bijma, P., and M. J. Wade, 2008 The joint effects of kin, multilevel selection and indirect genetic effects on response to genetic selection. Journal of Evolutionary Biology **21:** 1175-1188.

Bishop, S. C., and C. A. Morris, 2007 Genetics of disease resistance in sheep and goats. Small Ruminant Research **70:** 48-59.

Bishop, S. C., and M. J. Stear, 1997 Modelling responses to selection for resistance to gastro-intestinal parasites in sheep. Animal Science **64:** 469-478.

Bishop, S. C., and J. A. Woolliams, 2010 On the Genetic Interpretation of Disease Data. PLoS ONE **5:** e8940.

Breslow, N. E., and X. H. Lin, 1995 BIAS CORRECTION IN GENERALIZED LINEAR MIXED MODELS WITH A SINGLE-COMPONENT OF DISPERSION. Biometrika **82:** 81-91.

Charleston, B., B. M. Bankowski, S. Gubbins, M. E. Chase-Topping, D. Schley *et al.*, 2011 Relationship Between Clinical Signs and Transmission of an Infectious Disease and the Implications for Control. Science **332:** 726-729.

Chen, J., F. C. Michel, Jr., S. Sreevatsan, M. Morrison and Z. Yu, 2010 Occurrence and Persistence of Erythromycin Resistance Genes (erm) and Tetracycline Resistance Genes (tet) in Waste Treatment Systems on Swine Farms. Microbial Ecology **60:** 479-486.

Dawood, F. S., S. Jain, L. Finelli, M. W. Shaw, S. Lindstrom *et al.*, 2009 Emergence of a Novel Swine-Origin Influenza A (H1N1) Virus in Humans Novel Swine-Origin Influenza A (H1N1) Virus Investigation Team. New England Journal of Medicine **360:** 2605-2615.

Demeler, J., U. Kuettler, A. El-Abdellati, K. Stafford, A. Rydzik *et al.*, 2010 Standardization of the larval migration inhibition test for the detection of resistance to ivermectin in gastro intestinal nematodes of ruminants. Veterinary Parasitology **174:** 58-64.

Doeschl-Wilson, A. B., R. Davidson, J. Conington, T. Roughsedge, M. R. Hutchings *et al.*, 2011 Implications of Host Genetic Variation on the Risk and Prevalence of Infectious Diseases Transmitted Through the Environment. Genetics **188:** 683-693.

Ducrocq, V., J. Sölkner and G. Mészáros, 2010 Survival Kit v6 – a Software Package for Survival Analysis, pp. in *9th World Congress on Genetics Applied to Livestock Production*, Leipzig, Germany.

Ellen, E. D., J. Visscher, J. a. M. Van Arendonk and P. Bijma, 2008 Survival of laying hens: Genetic parameters for direct and associative effects in three purebred layer lines. Poultry Science **87:** 233-239.

Falconer, D. S., 1965 Maternal effects and selection response. Proc Int Congr Genet **3:** 763-774.

Falconer, D. S., and T. F. C. Mackay (Editors), 1996 *Introduction to Quantitative Genetics*. Pearson Education Limited, Harlow.

Gilmour, A. R., B. J. Gogel, B. R. Cullis and R. Thompson (Editors), 2006 *ASReml User Guide Release 2.0*. VSN International Ltd, Hemel Hempstead, UK.

Gimeno, I. M., 2008 Marek's disease vaccines: A solution for today but a worry for tomorrow? Vaccine **26:** C31-C41.

Griffing, B., 1967 Selection in reference to biological groups. I. Individual and group selection applied to populations of unordered groups Australian Journal of Biological Sciences **20:** 127-&.

Griffing, B., 1968a SELECTION IN REFERENCE TO BIOLOGICAL GROUPS .2. CONSEQUENCES OF SELECTION IN GROUPS OF ONE SIZE WHEN EVALUATED IN GROUPS OF A DIFFERENT SIZE. Australian Journal of Biological Sciences **21:** 1163-&.

Griffing, B., 1968b SELECTION IN REFERENCE TO BIOLOGICAL GROUPS .3. GENERALIZED RESULTS OF INDIVIDUAL AND GROUP SELECTION IN TERMS OF PARENT-OFFSPRING COVARIANCES. Australian Journal of Biological Sciences **21:** 1171-&.

Griffing, B., 1969 SELECTION IN REFERENCE TO BIOLOGICAL GROUPS .4. APPLICATION OF SELECTION INDEX THEORY. Australian Journal of Biological Sciences **22:** 131-&.

Griffing, B., 1976a Selection in reference to biological groups. 4. Use of extreme forms of nonrandom groups to increase selection efficiency Genetics **82:** 723-731.

Griffing, B., 1976b SELECTION IN REFERENCE TO BIOLOGICAL GROUPS .5. ANALYSIS OF FULL-SIB GROUPS. Genetics **82:** 703-722.

Griffing, B., 1981a A Theory of Natural-Selection Incorporating Interaction Among Individuals .1. the Modeling Process. Journal of Theoretical Biology **89:** 635-658.

Griffing, B., 1981b A Theory of Natural-Selection Incorporating Interaction Among Individuals .2. Use of Related Groups. Journal of Theoretical Biology **89:** 659-677.

Griffing, B., 1981c A Theory of Natural-Selection Incorporating Interaction Among Individuals .3. Use of Random Groups of Inbred Individuals. Journal of Theoretical Biology **89:** 679-690.

Griffing, B., 1981d A Theory of Natural-Selection Incorporating Interaction Among Individuals .4. Use of Related Groups of Inbred Individuals. Journal of Theoretical Biology **89:** 691-710.

Griffing, B., 1982a A Theory of Natural-Selection Incorporating Interaction Among Individuals .5. Use of Random Synchronized Groups. Journal of Theoretical Biology **94:** 709-728.

Griffing, B., 1982b A Theory of Natural-Selection Incorporating Interaction Among Individuals .6. Use of Nonrandom Synchronized Groups. Journal of Theoretical Biology **94:** 729-741.

Griffing, B., 1982c A Theory of Natural-Selection Incorporating Interaction Among Individuals .7. Use of Groups Consisting of One Sire and One Dam. Journal of Theoretical Biology **94:** 951-966.

Griffing, B., 1982d A Theory of Natural-Selection Incorporating Interaction Among Individuals .8. Use of Groups Consisting of A Sire and Several Dams. Journal of Theoretical Biology **94:** 967-983.

Heringstad, B., G. Klemetsdal and J. Ruane, 2000 Selection for mastitis resistance in dairy cattle: A review with focus on the situation in the Nordic countries. Livestock Production Science **64:** 95-106.

Heringstad, B., R. Rekaya, D. Gianola, G. Klemetsdal and K. A. Weigel, 2003 Genetic change for clinical mastitis in Norwegian cattle: a threshold model analysis. Journal of Dairy Science **86:** 369-375.

Jenkins, H. E., R. Woodroffe and C. A. Donnelly, 2010 The Duration of the Effects of Repeated Widespread Badger Culling on Cattle Tuberculosis Following the Cessation of Culling. Plos One **5**.

Keeling, M. J., and P. Rohani, 2008 *Modelling Infectious Diseases in Humans and Animals*. Princeton University Press, Princeton

Khuri, A. I., B. Mukherjee, B. K. Sinha and M. Ghosh, 2006 Design issues for generalized linear models: A review. Statistical Science **21:** 376-399.

Kimman, T. G., L. A. Cornelissen, R. J. Moormann, J. M. J. Rebel and N. Stochofe-Zurwieden, 2009 Challenges for porcine reproductive and respiratory syndrome virus (PRRSV) vaccinology. Vaccine **27:** 3704-3718.

Kirkpatrick, M., and R. Lande, 1989 The Evolution of Maternal Characters. Evolution **43:** 485-503.

Krebs, J. R., R. M. Anderson, T. Clutton-Brock, C. A. Donnelly, S. Frost *et al.*, 1998 Policy: Biomedicine - Badgers and bovine TB: Conflicts between conservation and health. Science **279:** 817-818.

Lande, R., and M. Kirkpatrick, 1990 SELECTION RESPONSE IN TRAITS WITH MATERNAL INHERITANCE. Genetical Research **55:** 189-197.

Lloyd-Smith, J. O., S. J. Schreiber and W. M. Getz, 2006 Moving beyond averages: Individual-level variation in disease transmission, pp. 235-258 in *Mathematical Studies on Human Disease Dynamics: Emerging Paradigms and Challenges*, edited by A. B. GUMEL. American Mathematical Society, Providence.

Lloyd-Smith, J. O., S. J. Schreiber, P. E. Kopp and W. M. Getz, 2005 Superspreading and the effect of individual variation on disease emergence. Nature **438:** 355-359.

Luo, M. F., P. J. Boettcher, L. R. Schaeffer and J. C. M. Dekkers, 2001 Bayesian inference for categorical traits with an application to variance component estimation. Journal of Dairy Science **84:** 694-704.

Mackenzie, K., and S. C. Bishop, 2001 Developing stochastic epidemiological models to quantify the dynamics of infectious diseases in domestic livestock. Journal of Animal Science **79:** 2047-2056.

Mcglothlin, J. W., and E. D. Brodie, 2009 How to Measure Indirect Genetic Effects: the Congruence of Trait-Based and Variance-Partitioning Approaches. Evolution **63:** 1785-1795.

Misztal, I., D. Gianola and J. L. Foulley, 1989 COMPUTING ASPECTS OF A NONLINEAR METHOD OF SIRE EVALUATION FOR CATEGORICAL-DATA. Journal of Dairy Science **72:** 1557-1568.

Moore, A. J., E. D. Brodie and J. B. Wolf, 1997 Interacting phenotypes and the evolutionary process .1. Direct and indirect genetic effects of social interactions. Evolution **51:** 1352-1362.

Muir, W. M., 2005 Incorporation of competitive effects in forest tree or animal breeding programs. Genetics **170:** 1247-1259.

Nath, M., J. A. Woolliams and S. C. Bishop, 2008 Assessment of the dynamics of microparasite infections in genetically homogeneous and heterogeneous populations using a stochastic epidemic model. Journal of Animal Science **86:** 1747-1757.

Ødegård, J., P. Madsen, R. Labouriau, B. Gjerde and T. H. E. Meuwissen, 2011 A sequential threshold cure model for genetic analysis of time-to-event data. Journal of Animal Science **89:** 943-950.

Ødegård, J., T. H. E. Meuwissen, B. Heringstad and P. Madsen, 2010a A simple algorithm to estimate genetic variance in an animal threshold model using Bayesian inference. Genetics Selection Evolution **42:** 1-7.

Ødegård, J., A.-I. Sommer and A. K. Praebel, 2010b Heritability of resistance to viral nervous necrosis in Atlantic cod (Gadus morhua L). Aquaculture **300:** 59-64.

Rauw, W. M., E. Kanis, E. N. Noordhuizen-Stassen and F. J. Grommers, 1998 Undesirable side effects of selection for high production efficiency in farm animals: a review. Livestock Production Science **56:** 15-33.

Rodriguez, G., and N. Goldman, 2001 Improved estimation procedures for multilevel models with binary response: a case-study. Journal of the Royal Statistical Society Series a-Statistics in Society **164:** 339-355.

Salzberg, S. L., C. Kingsford, G. Cattoli, D. J. Spiro, D. A. Janies *et al.*, 2007 Genome analysis linking recent European and African influenza (H5N1) viruses. Emerging Infectious Diseases **13:** 713-718.

Shen, Z., F. Ning, W. Zhou, X. He, C. Lin *et al.*, 2004 Superspreading SARS Events,Beijing, 2003. Emerging Infectious Diseases **10:** 256-260.

Stear, M. J., K. Bairden, S. C. Bishop, J. Buitkamp, J. L. Duncan *et al.*, 1997 The genetic basis of resistance to Ostertagia circumcincta in lambs. Veterinary Journal **154:** 111-119.

Tsuruta, S., and I. Misztal, 2006 THRGIBBS1F90 for estimation of variance components with threshold and linear models. Journal of Animal Science **84:** 15-15.

Van Oirschot, J. T., 2001 Present and future of veterinary viral vaccinology: A review. Veterinary Quarterly **23:** 100-108.

Vernon, M. C., and M. J. Keeling, 2012 Impact of regulatory perturbations to disease spread through cattle movements in Great Britain. Preventive Veterinary Medicine **105:** 110-117.

Wolf, J. B., E. D. Brodie, J. M. Cheverud, A. J. Moore and M. J. Wade, 1998 Evolutionary consequences of indirect genetic effects. Trends in Ecology & Evolution **13:** 64-69.

Wolf, J. B., E. D. Brodie and A. J. Moore, 1999 Interacting phenotypes and the evolutionary process. II. Selection resulting from social interactions. American Naturalist **153:** 254-266.

Chapter 2. Indirect Genetic Effects and the spread of infectious disease: are we capturing the full heritable variation underlying disease prevalence?

The following chapter has been published in PLOS One at the following URL: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0039551

Infectious diseases in livestock constitute a major threat to the sustainability of livestock production. Moreover, the need to contain epidemics has been further emphasized by the threat of transmission to other species – in particular humans - as illustrated in the recent swine flu epidemic (Dawood *et al.* 2009). Reducing disease prevalence through selection for host resistance offers a desirable alternative to chemical treatment which is a potential environmental concern due to run-off, and sometimes only offers limited protection due to pathogen resistance (Chen *et al.* 2010; Demeler *et al.* 2010). However, control of infectious diseases through selection has proven difficult as genetic analyses of resistance to infectious disease from field data tend to report low heritabilities (Bishop and Woolliams 2010). But is this a reflection of true genetic variance?

Current genetic analyses of disease data tend to focus on individual susceptibility to infectious disease, ignoring information from group members. However, using a stochastic epidemiological model, Nath et al. (2004) identified the transmission rate, latent period and recovery period as critical parameters for the risk and severity of infectious disease. In other terms, Nath et al. (2004) identified the impact that individuals have on each other as critical parameters for the risk and severity of infectious disease. Moreover, evolutionary theory would suggest that more genetic variation may be found in an individual's impact on its group mates than in

susceptibility. Since an individual's susceptibility is a component of its own fitness, natural selection works to exhaust heritable variation in susceptibility. An individual's impact on its group mates, in contrast, is not a component of its fitness, and may therefore accumulate greater heritable variation (Denison *et al.* 2003). As demonstrated by Van Dyken et al. (2011) this would occur even when kin-selection is acting, as populations in kin selection-mutation balance contain a stable frequency of 'cheaters'. In the context of disease, 'cheaters' correspond to hosts with increased shedding of infectious pathogens which have no damage to their own fitness but a potentially high cost to the herd. For example, assuming that animals with a greater parasite burden will also shed more, Raberg et al. (2007) found genetic variation in anaemia and weight loss corresponding to increasing parasite burden of rodent malaria in laboratory mice. These arguments suggest that there is an opportunity in capturing genetic variation in host infectivity, which is the propensity of transmitting infection upon contact with a susceptible individual. Especially as there is abundant evidence that heterogeneity in infectivity can profoundly impact upon disease prevalence in the population, with super-shedders being an extreme example (Doeschl-Wilson *et al.* 2011; Lloyd-Smith *et al.* 2005; Woolhouse *et al.* 1997; Yates *et al.* 2006).

Over the last forty years, the theory of Indirect Genetic Effects (IGE) has been developed to investigate the impact of interactions among individuals on the expression and evolution of traits (Bijma *et al.* 2007; Griffing 1967; Moore *et al.* 1997; Muir 2005). An indirect genetic effect, also known as an associative or social genetic effect, is a heritable effect of an individual on the trait value of another individual (Griffing 1967). Indeed, if an individual's trait value is affected by the genotypes of its population members (indirect genetic effect), then response to selection will be affected by these IGEs. It has been shown both theoretically (Bijma *et al.* 2007; Griffing 1967; 1976a) and experimentally (Muir and Craig 1998) that IGEs can drastically affect the rate and direction of response to selection. In this context, host infectivity can be regarded as an indirect effect to disease status. Thus an individual's disease status and infectious disease prevalence in a population is

likely to be affected by host genetic variation in both susceptibility and infectivity. To date, however, no work has been published examining the prospects of IGE models for infectious diseases, suggesting that part of the heritable variation underlying disease prevalence is overlooked.

Genetic variation in infectivity is difficult to measure directly and may need to be inferred from more readily available information such as binary disease data (infected/non-infected). Our hypothesis is that current genetic models applied to binary disease data do not capture the full genetic variation underlying disease prevalence and that a model including IGEs is more appropriate. This study, therefore, examines to what extent genetic variance in infectivity/susceptibility is captured by a conventional model versus an IGE model in populations with simulated genetic variation in infectivity, and whether selection on breeding values estimated with IGE models offer greater potential for reducing disease prevalence. In order to address this question, we modelled disease progression in populations with different genetic architectures for infectivity/susceptibility and estimated the genetic variance in the simulated binary disease data with a conventional animal model and a model including IGEs. Finally, we evaluated selection response in susceptibility and infectivity, and its impact on future disease risk, using the estimated breeding values (EBV) derived from both models.

2.1 Methods

2.1.1 The epidemiological model

An epidemic was simulated to describe disease progression in the population and provide, as output, the disease status of each individual at given time points. To avoid overburdening the results with unnecessary complexity we chose a simple compartmental stochastic SIR model of disease spread modified from Keeling and

Rohani (2008). In an SIR model it is assumed that individuals start as being susceptible (S) but may then become infected (I), upon contact with an infected individual, eventually recover (R) and are then no longer susceptible. The speed of transition between the epidemiological compartments S, I, and R is determined by the transmission parameter β (S->I) and by the recovery rate γ (I->R). It was also assumed that infected individuals become immediately infectious. The contact between individuals within a group was constant and uniform (contact rate = 1) and no transmission was allowed between groups.

To allow for individual genetic variation in the epidemiological parameters β and γ, each individual j was assigned its own level of susceptibility g_j, infectivity f_j and speed of recovery γ_j. The pairwise transmission parameter β_{jk} was then defined as

$$\beta_{jk} = -ln(1 - X_{g,j} g_j X_{f,k} f_k). \quad (2.1)$$

We refer to Appendix 1 for the derivation of equation (2.1). For ease of reading a comprehensive list of symbols and notation is given in Table 2-1. Thus β_{jk} is a function of the product of the susceptibility g of individual j and the infectivity f of individual k. To reflect whether susceptibility is expressed by individual j, it is scaled by $X_{g,j}$ which equals one if j is susceptible and zero otherwise. Similarly, infectivity is scaled by $X_{f,k}$ which equals one if k is infected and zero otherwise. For simplicity, it was assumed that infectivity and susceptibility are independent, and that all individual speeds of recovery γ_j were equal to a constant $\gamma = 0.1$ if the individual was infected and zero otherwise.

The epidemic was simulated as a Poisson process, i.e. as a series of random independent events occurring at given average rates in continuous time. In this model the possible events were infection of a susceptible individual and recovery of an infected individual. The average infection rate r_I within a group was estimated as the sum of the pairwise transmission parameters β_{jk} of the group members and the average recovery rate r_R as the sum of the individual speeds of recovery γ_j.

The simulated epidemic was started by a single randomly chosen infected individual within each group of size *n* in an otherwise naïve population. The time to the next event (inter-event times) and the corresponding event type (infection of a susceptible individual or recovery of an infected individual) were then estimated using Gillespie's direct algorithm (Gillespie 1977) which is a commonly used algorithm in stochastic epidemiological models (Keeling and Ross 2008). Specifically, the probability density function describing the time between events in a Poisson process is the exponential distribution with scale parameter equal to the average rate at which events occur. Thus inter-event times for each group were sampled from an exponential distribution with parameter $r = r_I + r_R$. In other words, the time between each event was estimated as $-\ln(x_1)/r$ where $x_1 \sim U(0,1)$. The specific event type *v* (i.e. infection or recovery) which then occurs was obtained by drawing a random variate from a discrete distribution with probability $p(v) = r_v / r$. Hence, the event was an infection if $x_2 < r_I / r$ where $x_2 \sim U(0,1)$ and a recovery otherwise. The individual involved in each event was then chosen randomly weighted by the individuals' susceptibility or recovery rate. No transmission was assumed between groups.

Table 2-1 Symbols and notations

g_j	Susceptibility of individual *j*
f_j	Infectivity of individual *j*
γ_j	Speed of recovery of individual *j*
γ	Speed of recovery constant
β_{jk}	Pairwise transmission parameter from individual *k* to individual *j*
r_I	Average rate of infection
r_R	Average rate of recovery
x_1, x_2	Random variates
U(0,1)	Uniform distribution between zero and one
N	Population size
n	Group size
μ	Fixed mean of susceptibility and infectivity
G1	Effect of allele with small effect on susceptibility, in bi-allelic architecture
G2	Effect of allele with large effect on susceptibility, in bi-allelic architecture
F1	Effect of allele with small effect on infectivity, in bi-allelic architecture
F2	Effect of allele with large effect on infectivity, in bi-allelic architecture
MAF	Minor allele frequency (right-skewed distribution, applies to allele with large effect)
α	Allele substitution effect
$N(\mu,\sigma^2)$	Normal distribution with mean μ and variance σ^2
$\Gamma(a,\theta)$	Gamma distribution with shape *a* and scale θ
σ^2_A	Genetic variance from conventional model
σ^2_D	Direct genetic variance from IGE model
σ^2_S	Indirect genetic variance from IGE model
σ^2_e	Residual variance
b_1, b_2	Regression coefficients
$\bar{p}$	Mean number of infected group mates
EBV_A	Estimated Breeding Values from the conventional model
EBV_D	Estimated Breeding Values for the direct effect from the IGE model
EBV_S	Estimated Breeding Values for the indirect effect from the IGE model
I_x	Index of Estimated Breeding Values

R_0	Basic reproduction number: expected number of secondary infections caused by an individual in its lifetime.

2.1.2 Simulated populations

In order to ensure a high power to detect genetic variation, large populations with a relatively large family size and a family structure following dairy cattle, for example, were simulated. In particular, populations of size N=100,000 were created with a paternal half-sib structure and no full sibs. All parents were assumed to be unrelated. The half sib family size was 100 individuals. Similarly, in order to ensure a high power to detect genetic variation, each population was divided into 10,000 groups of size 10 chosen at random without reference to pedigree.

Breeding values for susceptibility and infectivity were assigned to the individuals in the parental generation using different distributions to account for different underlying genetic architectures. For the first architecture it was assumed that genetic variation in susceptibility was controlled by a single bi-allelic locus and genetic variation in infectivity by another bi-allelic locus. Both loci were assumed to segregate independently. This architecture was used to encompass diseases affected by a major gene. For example, Houston et al. (2010) found that a single quantitative trait locus (QTL) explained 98% of the additive genetic variation in susceptibility to infectious pancreatic necrosis (IPN) in salmon. For the second architecture it was assumed that genetic variation in these traits is influenced by many alleles conferring a continuous distribution of effect sizes (possibly stemming from several loci).

Parametric statistical analyses usually assume normality. However, as shown by Lloyd-Smith et al. (2005), the distribution of infectivity is often right-skewed. Moreover, skewed distributions allow for larger variation when the distribution is confined to positive values. Both types of genetic architectures were, therefore, considered with either a symmetrical or a right-skewed frequency distribution. In all four combinations (two alleles - symmetric, two alleles - skewed, multiple alleles - symmetric, multiple alleles - skewed) mean susceptibility and infectivity were fixed

at $\mu = 0.22$, as different population means would lead to different prevalence profiles. Fixing the means does however imply that populations with different genetic architectures have different input variances, and may thus not be directly comparable. However the focus of the study is comparison of animal models *vs.* IGE models within a genetic architecture.

2.1.2.1 Two alleles genetic architecture

For the bi-allelic architecture, it was assumed that the locus influencing susceptibility has two alleles each inferring a value of *G1* or *G2* and the locus influencing infectivity has two alleles each inferring a value of *F1* or *F2*. We also assumed additivity of allelic effects without dominance and that the population is in Hardy-Weinberg equilibrium. In other words, the genetic values for susceptibility (or infectivity) in the parental population were sampled from a discrete distribution with three possible values *G1+G1* (or *F1+F1*), *G1+G2* (or *F1+F2*) and *G2+G2* (or *F2+F2*). The shape of the distribution was defined through the minor allele frequency (MAF) which applied to the allele with a large effect (*F2*, *G2*). The values corresponding to each of the alleles were then chosen such that the population mean and the allele substitution effect α were kept constant. The same parameters were used for both infectivity and susceptibility to facilitate comparison of estimated genetic parameters. Table 2-2 shows the parameter values for the bi-allelic genetic architecture. The offspring's breeding values were then generated by randomly allocating dams to sires and randomly choosing one allele from each parent.

2.1.2.2 Multiple alleles genetic architecture

For the multiple allele architecture, it was assumed that there would be sufficient alleles contributing to the additive genetic values of susceptibility and infectivity in the parental population to be adequately approximated by a continuous probability density function.

For the symmetric frequency distribution, the breeding values for the parental population were sampled from the normal distribution N(μ, σ^2). The parameter values were taken as μ=0.22 (i.e. the same as for the bi-allelic architecture) and σ^2=0.005 to avoid frequent negative values of susceptibility/infectivity (Table 2-2). If a negative value was sampled, it was discarded and re-sampled. Each offspring was allocated a breeding value equal to the mean of its parents plus a Mendelian sampling term.

For the skewed frequency distribution, the breeding values of the parental population for susceptibility and infectivity were assumed to be distributed according to the gamma distribution $\Gamma(a,\theta)$. It is not possible to represent Mendelian inheritance by adding a Mendelian sampling term with a gamma distribution, however, as the offspring generation would no longer follow the same distribution as the parental generation. It was therefore assumed that the parental breeding values stem from ten additive loci with a large number of alleles each, whose effect follow the gamma distribution $\Gamma(a/20,\theta)$. The offspring were then randomly assigned one effect from each parent for each locus. The breeding values of the offspring are therefore distributed following $\Gamma(a,\theta)$, given that for any n Gamma distributed variates $G_i \sim \Gamma(k_i,\xi)$ the distribution of their sum is given by $\sum_{i=1}^{n} G_i \sim \Gamma(\sum_{i=1}^{n} k_i, \xi)$. Note that if the shape parameter $\sum k_i$ became large the Gamma distribution would approach the Normal distribution in accordance to the central limit theorem. Specifically, the parameters were taken as a=1.1 and θ=0.2 such that the mean $a\theta=\mu$=0.22, i.e. the same as for the bi-allelic architecture, the variance $a\theta^2$=0.044 (Table 2-2) and the distribution is right-skewed (skewness $2/\sqrt{a}$=1.9).

For all populations, it was assumed that susceptibility and infectivity are fully heritable and that the outcome, i.e. whether an individual becomes infected or not, depends on both the genetics and environment. The environmental contribution to the phenotypic variance was represented through the stochastic events (infection, recovery) in the epidemiological model. Thus, the model assumes genetic predisposition whilst maintaining full environmental stochasticity of the epidemics. Moreover, adding additional environmental noise would not provide further useful

information to this study and would make it harder to interpret the results. Each architecture was run with variation introduced in susceptibility only, infectivity only, both or neither. When no variation in susceptibility/infectivity was introduced, all individuals were given a fixed breeding value of $\mu = 0.22$ for that underlying trait. As each simulated population is divided into 10,000 groups, i.e. 10,000 independent epidemics, each simulation was replicated ten times.

Table 2-2 Parameters for Breeding Values generation

M.A.F. applied to the alleles with a large effect (F2, G2).

		M.A.F.	Allele values		α	Population mean μ	Variance
Distribution			F1,G1	F2,G2			
Two alleles	Symmetric	0.5	0.02	0.2	0.18	0.22	0.0162
	Skewed	0.2	0.074	0.254	0.18	0.22	0.0104
Multiple alleles	Symmetric	-	-	-	-	0.22	0.0049
	Skewed	-	-	-	-	0.22	0.0440

2.1.3 Estimating genetic variance

Genetic variation between individuals was estimated from binary records which were obtained by recording the disease state of simulated individuals. The binary disease trait, disease presence, was one if an individual had become infected prior to a considered time-point and zero otherwise. The data were analysed at the same time point for all groups, which was the time at which 50% of individuals would have become infected in a homogeneous population with the same mean values for the input parameters. All analyses were carried out using ASReml (Gilmour *et al.* 2006).

To reflect current practise, genetic variance in disease presence was first estimated with a mixed model including a single genetic variance. In order to be in line with the IGE model, this was achieved with an animal model for disease presence y observed in offspring j of sire i,

$$y_{jh} \sim \text{mean} + (\text{animal effect})_j + e_{jh}. \quad (2.2)$$

The group effect is absorbed by allowing for a correlation ρ between the residuals of group members, this is equivalent to fitting a random group effect (Bergsma *et al.* 2008). The animal variance is denoted as σ_A^2. Hereafter this model is referred to as the conventional model.

To estimate the genetic variance in the indirect effect, the data were analysed using the model developed by Muir (2005). Thus for disease presence *y* observed in offspring *j* with this individual living in group *h* of size *n* with group mates m,

$$y_{jh} \sim \text{mean} + (\text{direct effect})_j + \sum_{m=1}^{n-1} (\text{associative effect})_{mh} + e_{jh}. \quad (2.3)$$

Similarly to the conventional model, the group effect is absorbed by allowing for a correlation between the residuals of group members (Bergsma *et al.* 2008). Note that this model does not take account of the disease status of individuals *j* and their group members *m*, in other words, it is assumed that all individuals express the direct effect (susceptibility) and the indirect effect (infectivity) at all times, regardless of their infection status. The variances of the direct and indirect genetic effects are denoted σ_D^2 and σ_S^2 respectively. Hereafter this model is referred to as the Indirect Genetic Effects (IGE) model.

2.1.4 Association between variation in susceptibility/infectivity and variation in binary disease presence

In order to assess to what extent the available genetic variation is being captured by the different statistical models, an estimate of expected output variance as a function of the input variance in infectivity/susceptibility is required. Following Dempster and Lerner (1950) a linear relationship was assumed between input and output traits to provide an approximation. In particular, it was assumed that there is a linear

relationship between disease presence in an individual *j* and that individual's susceptibility g_j and the sum of the infectivities *f* of the *p* infected group mates of that individual,

$$y_{jh} \sim \text{mean} + b_1 g_j + b_2 \sum_{m=1}^{p} f_{mh} + e_{jh}. \qquad (2.4)$$

The regression mean and coefficients b_1 and b_2, were estimated using this linear model in the statistical package R (Ihaka and Gentleman 1996) with the known input (i.e. true *f* and *g* values) and output (*y*) data from the simulations. Hence the model in equation (2.4) was used to estimate the true linear effects of infectivity and susceptibility to the observed binary disease presence.

The number of group mates that have been infected (p) is a variable which depends on the group *h* and status of individual *j*. Indeed, if in a given group x individuals have been infected, individual *j* will have *x* group mates which have been infected if it is susceptible and *x-1* if it is one of the infected individuals. The variance of disease presence σ^2 may therefore be expressed as follows (cf. derivation in Appendix 2):

$$\sigma^2 = b_1^2\sigma_g^2 + b_2^2\bar{p}\sigma_f^2 + b_2^2\bar{f}^2\sigma_p^2 + \sigma_e^2. \quad (2.5)$$

This expression can be compared with the estimated variance of disease presence $\hat{\sigma}^2$ that is obtained from the IGE model in equation (2.3),

$$\hat{\sigma}^2 = \sigma_D^2 + (n-1)\sigma_S^2 + \sigma_{e'}^2 + (n-1)\rho\sigma_{e'}^2. \qquad (2.6)$$

The first term in equation (2.5) is a function of the input variance in susceptibility σ_g^2 and should be approximately comparable to σ_D^2 from the IGE model and to σ_A^2 from the conventional model. The second term in equation (2.5) is a function of the input variance in infectivity σ_f^2 and mean number of infected group mates $\bar{p}$ over all

groups, and should be approximately comparable to $(n-1)\sigma_S^2$, i.e. the second term in equation (2.6). The third term is a function of the squared input mean infectivity $\bar{f}^2$ and the variance in number of infected group mates σ_p^2. It is not directly comparable with any ASReml output as this term includes both between group variation and interaction between infectivity and susceptibility. Note that the expression of infectivity depends on the individual being infected, which in turn depends on the individuals' own susceptibility, and σ_p^2 can be said to be the variation in numbers of individuals expressing infectivity. The interdependence in this model between infectivity and susceptibility is likely to be partly captured through a non-zero covariance estimate between direct and indirect genetic effects in ASReml (Gilmour *et al.* 2006).

2.1.5 Estimated response to selection

In order to estimate response to selection based on Estimated Breeding Values (EBVs) derived from the conventional and IGE models, the impact of selection on true mean susceptibility/infectivity was examined. Here the population mean susceptibility/infectivity was compared to the mean susceptibility/infectivity after selection of 10% of the individuals with the lowest EBVs obtained from each model. For the conventional model, selection used the only available EBV (EBV_A). For the IGE model, selection was based on the EBVs for direct (EBV_D) and indirect (EBV_S) genetic effect separately as well as for the index $I_x = EBV_D + (n-1)\,\bar{p}EBV_S$. The weight of the index was selected to take the mean level of exposure i.e. $(n-1)\,\bar{p}$ into account.

To quantify response to selection in terms of risk and severity of the epidemic, the basic reproduction number R_0 was estimated for the whole population and for each selected subpopulation using the true values of susceptibility and infectivity. R_0 is the mean number of secondary infections an infected individual will cause in its lifetime and is commonly used as a measure of disease risk and severity in epidemiology

(Anderson and May 2006). By definition, an epidemic will die out if $R_0<1$. Following a SIR model for a closed population, $R_0 = \beta S_0 / \gamma$, with $S_0=(n-1)$ being the initial number of susceptible individuals in a group (Keeling and Rohani 2008). Incorporating equation (2.1) and taking a Taylor series expansion we obtain,

$$R_0 = \left(-\ln(1-\bar{g}\bar{f})S_0\right)/\gamma. \qquad (2.7)$$

The symmetry of susceptibility and infectivity in equation (2.7) implies that a decrease in mean susceptibility or infectivity will decrease mean R_0 equally (cf. Figure S1, Appendix 3).

2.2 Results

2.2.1 Estimated genetic variance in disease presence using a conventional model

The estimated variances in disease presence obtained for each population using a conventional model, along with the mean presence over all groups in all replicates, are displayed in Table 2-3. Overall the variance estimates depend on input variance and on mean presence at time of evaluation. As input parameters were the same for susceptibility and infectivity, variance estimates along the rows of Table 2-3, where mean presence is the same, are directly comparable. Note that values in rows are not directly comparable, on the other hand, across columns with different mean presence.

The results indicate that, if there is variation in infectivity only, the conventional model fails to pick up the heritable variation in binary disease presence present in the data. Only in the populations with a skewed multiple allele genetic architecture (i.e. large variance in infectivity) a small amount of genetic variation is captured when there is variation in infectivity only. However, the resulting variance estimate was

only 3.5% of that compared to populations with the same variance introduced in susceptibility.

Table 2-3 Estimated genetic variance in disease presence (binary) using a conventional animal model

All parameters as in Table 2-2. 10000 groups of size 10, values ± empirical between replicate standard deviation, '#' means not significantly different from zero (P > 0.05), values scaled by 10^3.

Distribution			Variation introduced in: None	Infectivity	Susceptibility	Both
Two alleles	Symmetric	Variance	0.32#±0.08	0.63#±0.09	25.35±0.27	18.74±0.45
		Mean presence	0.56	0.50	0.51	0.46
	Skewed	Variance	0.32#±0.08	0.37#±0.10	8.28±0.14	7.96±0.13
		Mean presence	0.56	0.53	0.53	0.51
Multiple alleles	Symmetric	Variance	0.13#±0.04	0.09#±0.08	0.12#±0.09	0.10#±0.03
		Mean presence	0.51	0.48	0.49	0.50
	Skewed	Variance	0.24#±0.08	0.74±0.09	31.02±0.53	18.56±0.45
		Mean presence	0.51	0.42	0.42	0.35

2.2.2 Estimated genetic variances using an IGE model

Given that the prevalence profiles were similar between genetic architectures (cf. Figures S2 & S3, Appendix 3) and the skewed multiple alleles population had the largest input variance, analyses using the IGE model were only performed on populations with skewed distributions for susceptibility and infectivity.

The genetic parameters obtained by analysis with the IGE model along with relevant statistics are displayed in Tables 2-4 & 2-5. Note that following equation (2.6) the contribution of the indirect genetic effect to the phenotypic variance is (n-1) times greater than the values in Tables 2-4 & 2-5. Variance in infectivity is captured by the σ_S^2, in populations with both genetic architectures (cf. Tables 2-4 & 2-5). A log-likelihood test was performed to evaluate the statistical significance of the indirect genetic effects term. As would be expected, the indirect genetic effects term was

significant ($P < 0.05$) in populations with variation in infectivity (cf. Tables 2-4 & 2-5). The analysis of the skewed multiple alleles population also implies that there is a statistically significant positive genetic covariance between the direct and the indirect effect when there is variance in susceptibility (cf. Table 2-5), despite susceptibility and infectivity being independent in our simulation. This is probably due to the fact that the model fitted assumes constant expression of effects by all group members whereas an individual will only express infectivity if infected, which will depend on the individual's susceptibility. Note that the values in Tables 2-4 & 2-5 were obtained from the same data as those in Table 2-3, so the values in Table 2-3 can be compared to those in Tables 2-4 & 2-5.

Table 2-4 Estimated genetic variance in disease presence (binary), in populations with a skewed bi-allelic genetic architecture underlying susceptibility/infectivity, using the Indirect Genetic Effects model

Values scaled by 10^3, values ± empirical between replicate standard deviation, '#' means not significantly different from zero (P>0.05). Values along the rows are directly comparable to each other where mean presence is the same. Estimates averaged over ten iterations. Parameter values as in Table 2-2, 10000 groups of size 10. The log-likelihood P-value refers to the significance of the indirect genetic effect.

	Variation introduced in:			
Estimated genetic variance/ covariance in	None	Infectivity	Susceptibility	Both
Direct effect σ^2_D	$0.32^{\#}$±0.09	$0.22^{\#}$±0.11	9.19±0.30	8.63±0.16
Indirect effect σ^2_S	$0.14^{\#}$±0.04	0.51±0.04	$0.16^{\#}$±0.03	0.43±0.05
Direct/indirect effect σ_{DS}	$0.06^{\#}$±0.04	$0.08^{\#}$±0.08	$0.45^{\#}$±0.14	$0.59^{\#}$±0.13
Log likelihood test P-value	0.4	$0.3*10^{-2}$	0.5	0.04
Mean presence	0.56	0.53	0.53	0.51

Table 2-5 Estimated genetic variance in disease presence (binary), in populations with a skewed multiple alleles genetic architecture underlying susceptibility/infectivity, using the Indirect Genetic Effects model

Values scaled by10^3, values ± empirical between replicate standard deviation, '#' means not significantly different from zero (P>0.05). Values along the rows are directly comparable to each other where mean presence is the same. Estimates averaged over ten replicates. Parameters as in Table 2-2, 10000 groups of size 10. The log-likelihood P-value refers to the significance of the indirect genetic effect.

	Variation introduced in:			
Estimated genetic variance/ covariance in:	None	Infectivity	Susceptibility	Both
Direct effect σ_D^2	0.26#±0.09	0.36#±0.11	28.07±1.97	19.55±0.47
Indirect effect σ_S^2	0.16#±0.03	1.00±0.09	0.11#±0.04	0.43±0.04
Direct and indirect effect σ_{DS}	0.08#±0.05	0.13#±0.09	1.05±0.29	0.86±0.09
Log likelihood test P-value	0.5	$0.2*10^{-5}$	0.5	$0.3*10^{-2}$
Mean presence	0.51	0.42	0.42	0.35

2.2.3 Comparison of input and estimated variances

Input variance in susceptibility and infectivity and estimated variances were brought to a comparable scale using equations (2.5) and (2.6) and are displayed in Table 2-6. From Table 2-6 it is evident that the first term in equation (2.5), σ_D^2, and σ_A^2 are approximately similar. However, the second term in equation (2.5) appears to be consistently larger than (n-1)σ_S^2, suggesting that the IGE model underestimates variation in infectivity. This could be due to the fact that the IGE model assumes constant expression of infectivity by all group-members, whereas in equation (2.5) expression of infectivity is limited to infected individuals. In this way the indirect effect is distributed between (n-1) individuals, in the genetic analysis with the IGE model, compared to $\bar{p}$ in equation (2.5) with $\bar{p} \le n-1$. The discrepancy in these variance estimates suggests that there is some scope for improvement.

Table 2-6 A comparison of expected and observed variance components for the skewed 'multiple alleles' and 'two alleles' architectures when genetic variance is introduced INTO infectivity, or susceptibility, or both

Observed components are taken from results of analyses of data with either a conventional model (Eqn 2.2) or IGE model (Eqn 2.3), whilst expected components are obtained from the true simulated values and Eqn 5. '#' means not significantly different from zero (P>0.05), values scaled by 10^3.

Variation introduced in:		Expected: Susceptibility $b_1^2\sigma_g^2$	IGE: Direct σ_D^2	Conventional: σ_A^2	Expected: Infectivity $b_2^2\bar{p}\sigma_f^2$	IGE: Indirect $(n-1)\sigma_S^2$
Multiple Alleles	Infectivity	0.00	0.36#	0.74	15.46	9.04
	Susceptibility	36.46	28.07	31.02	0.00	0.99#
	Both	20.39	19.55	18.56	9.20	3.87
Two Alleles	Infectivity	0.00	0.22#	0.37#	6.34	4.59
	Susceptibility	8.86	9.19	8.60	0.00	1.44#
	Both	7.92	8.63	7.96	5.34	3.87

2.2.4 Impact of selection on mean susceptibility/infectivity and future disease risk

Mean susceptibility and infectivity, of the whole population and selected sub-populations, together with their respective average R_0 values are displayed in Table 2-7. In line with our previous results, selection on the breeding values derived from the conventional model or on EBV_D alone, only reduces mean susceptibility (cf. Table 2-7). Less predictably, however, selection on EBV_S reduced both mean infectivity and susceptibility (cf. Table 2-7). This may be due to expression of infectivity being dependent on being infected, which in turn depends on susceptibility as mentioned above. This suggests that, when status is not taken into account, selection targeting infectivity would indirectly also select for lower susceptibility. However, the resulting average R_0 values displayed in Table 2-7

suggest that an index with both direct and indirect breeding values would create the greatest impact for the reduction of disease in future generation.

Table 2-7 Mean susceptibility and infectivity following selection using the conventional animal model or the Indirect Genetic Effects model

Population with variation in both infectivity and susceptibility following a skewed multiple allele genetic architecture. 10000 groups of size 10. Proportion selected was 0.10. Values ± standard error when greater than 0.005.

Selection		Mean susceptibility	Mean infectivity	R_0
None		0.22	0.22	4.46
Conventional animal effect	EBV	0.10	0.22	1.99 ± 0.04
Direct effect	EBV_D	0.10	0.22	1.96 ± 0.04
Indirect effect	EBV_s	0.15	0.17	2.38 ± 0.11
Index	$Ix=EBV_D+\bar{p}$ (n-1) EBV_s	0.11	0.19	1.91 ± 0.03

2.3 Discussion

The hypothesis of this study was that low heritability estimates of disease traits may not reflect the true additive genetic variation inherent in a population, but rather a deficiency in the philosophy underpinning the models that are currently fitted. The aim of this study was therefore to assess whether it is possible to capture genetic variation in infectivity, when it is inherent in the data, with current statistical methods (animal/sire and IGE model). This was assessed for a variety of genetic architectures underlying susceptibility and infectivity. Our results show that, unlike a conventional model, which does not capture the variation in infectivity when it is present in the data, a model which takes indirect genetic effects (IGE) into account captures some, though not all, of the inherent genetic variation in infectivity. This implies that, failing to include IGEs when analysing disease data from field studies may result in substantial genetic variation being missed. For example had the QTL, explaining 98% of the additive genetic variation in susceptibility to pancreatic necrosis in salmon, found by Houston et al. (2010) affected infectivity rather than

susceptibility it would probably have been overlooked. Moreover, this additional genetic variance does not come at the expense of obtaining reliable estimates for genetic variance in susceptibility.

Our results show further that the ability of IGE models to detect genetic variance in infectivity can impact on the response to subsequent artificial selection. From the mean susceptibility/infectivity and R_0 values of the selected subsets of the population it is evident that, even with BVs estimated with the current IGE model based on binary data from a single time point, a greater impact on disease risk and severity could be achieved than when using BVs estimated with a conventional model. This is particularly true in populations with variation in infectivity only, as no selection would have been possible based on breeding values derived from a conventional model. At present, it is unknown whether infectivity harbours substantial genetic variation, or whether populations with genetic variation in infectivity only are common. This work, however, provides the first tools to address these questions.

Comparison with expected genetic variance from an alternative model using linear approximations suggests that there is still scope for improvement in applying IGE models to disease data. The apparent underestimation of genetic variance in infectivity may be due to the fact that the current methodology does not allow for status dependence. This could potentially cause an underestimation of the variance in infectivity as the IGE is attributed to all individuals in a group, when in reality it will have been expressed by only a subset of group members. Furthermore, our analysis revealed that the statistical model applied here is likely to yield a positive covariance estimate despite susceptibility and infectivity being independent. This is probably because expression of infectivity is state dependent and thus partly depends on the individual's susceptibility. Allowing for status dependency should therefore improve the accuracy of the estimated genetic parameters, thus suggesting that responses to selection may be greater than values presented here when methods are further improved.

The data of this study were generated using a standard epidemiological SIR model, assuming only host genetic variation in susceptibility and infectivity and full independence and heritability for both traits, in order to reduce unnecessary noise. Moreover, potential host-pathogen interactions were not considered. Although these assumptions may be representative for a variety of infectious diseases and populations, one would expect that the different sources of variances for diseases with more complex epidemiological patterns and in populations with more complex variance and co-variance structures would be more difficult to capture. This enhances the need for further investigations of IGE models with regards to requirements for data collection and experimental design for obtaining reliable genetic parameter estimates corresponding to host susceptibility and infectivity.

In addition to susceptibility and infectivity investigated here, there may be other sources of host genetic variation contributing to genetic variance of disease data and thus amenable for selection. For example, in addition to variation in infectivity, i.e. the propensity of individuals to infect others upon contact, genetic differences in transmission patterns may be caused by heritable variation in contact rate due to behavioural traits such as aggression or promiscuity. Previous studies have demonstrated (Bergsma *et al.* 2008) that IGE models are able to provide reliable estimates for these social interactions. Moreover, additional heritable variation in disease presence may come from genetic differences in recovery time among individuals, which affects their infective period. Analyses accounting for genetic differences in the length of the infective period may contribute to achieving greater response to selection, and emphasizes the scope for additional work in this area. We achieved a first step in understanding and extending the range of epidemiological parameters under potential host genetic influence that can be estimated with current quantitative genetic models. Further work is required to increase our understanding and improve the statistical models through the use of simulations and the application to field data.

Bishop and Woolliams (2010) have shown that accuracy of genetic parameters for disease data obtained from field studies depends largely on exposure, and thus on

time of measurement. Disease records obtained at a time corresponding to high disease prevalence are expected to give higher heritability estimates than disease records obtained at times when prevalence was low. It is expected that similar relationships also apply for the estimation of genetic parameters associated with indirect genetic effects. Further, Bijma (2010) has shown that substantial improvement in accuracy of indirect genetic variance components can be achieved by optimising group size and composition. Since group size has a strong effect on disease progression between individuals and thus on prevalence patterns (cf. Appendix 3, Figures S2 & S3), it is expected that much improvement in the estimation of indirect genetic effects could be obtained by choosing the correct combination of group size and time at which records are collected. This could be combined with groups composed of members of two families, which yields much better accuracy of estimated genetic parameters than groups composed at random, particularly when groups are large (Bijma 2010). Moreover, different weightings for the direct and indirect effects EBVs in the index might offer further improvements depending on the context.

One of the remaining challenges of analysing binary disease data with an IGE model is to establish the relationship between underlying susceptibility/infectivity and direct/indirect genetic effects. There are two standard ways of estimating genetic parameters from a binary trait, either using a linear mixed model, which treats the data as continuous and includes random factors, or a generalised linear model (GLM). The use of a GLM in combination with random factors (GLMM) is an area that is open to question. In fact, ASReml (Gilmour *et al.* 2006), the software used to fit the models in this paper, provides a warning not to use a GLM in combination with random factors. The relationship between the underlying traits, susceptibility and infectivity, and the observed trait disease presence is complex and stochastic. It is therefore questionable whether canonical link functions relating underlying parameters (e.g. susceptibility, infectivity) with the probability of observing an event (e.g. becoming infected), such as probit or logistic functions, would be appropriate in our case. In fact variance estimates obtained using a logistic model are not on the same scale as susceptibility and infectivity (cf. Appendix 4, Table S2). Moreover,

should we use non-standard distributions and link functions, further statistical issues would arise, e.g. decomposing the phenotypic variance into genetic and environmental components may no longer be valid. Hence there is no theoretical apparent benefit in applying specific link functions with a GLM. Moreover, variance estimates obtained with a logistic model are qualitatively the same as those obtained with the linear models (cf. Appendix 4, Table S2). Besides, selection on the EBVs obtained with a logistic model provided no better results with regards to R_0 (cf. Appendix 4, Table S3). We therefore decided to use a linear mixed model, which have been shown to provide estimates of genetic parameters of sufficient accuracy to generate selection response e.g. (Ramirez-Valverde *et al.* 2001; Vazquez *et al.* 2009).

Better understanding of the factors involved in indirect genetic effects to disease presence could open up further potential for disease control through selection. For example, it has been shown that, when IGEs occur, response to selection depends on the covariance between the direct and indirect genetic effects (Griffing 1967), which correspond to susceptibility and infectivity in our study. In this study we assumed that infectivity and susceptibility are independent. However, should they be positively correlated, the expected response to selection would be greater than indicated here. Doeschl-Wilson et al. (2008) demonstrate for gastro-intestinal parasitism in sheep that correlation between underlying disease traits can have profound impact on heritabilities of observable disease traits and thus on response to selection. Moreover, a recent study showed molecular evidence for a positive correlation between susceptibility and infectivity as the known immunosuppressant stress hormone norepinephrine was shown to cause increase shedding of Salmonella (Pullinger *et al.* 2010). It is therefore reasonable to believe that being able to estimate variance in indirect genetic effects for disease may open up new avenues for the control of infectious diseases through selection. In conclusion, this is the first work on the relevance of IGEs for the spread of infectious disease and it indicates that their relevance extends beyond behavioural interactions among individuals, which is the current focus of such research e.g. (Wilson *et al.* 2009).

2.4 References

Anderson, R. M., and R. M. May, 2006 *Infectious Diseases of Humans*. Oxford University Press, Oxford.

Bergsma, R., E. Kanis, E. F. Knol and P. Bijma, 2008 The contribution of social effects to heritable variation in finishing traits of domestic pigs (Sus scrofa). Genetics **178:** 1559-1570.

Bijma, P., 2010 Estimating Indirect Genetic Effects: Precision of Estimates and Optimum Designs. Genetics **186:** 1013-1028.

Bijma, P., W. M. Muir, E. D. Ellen, J. B. Wolf and J. a. M. Van Arendonk, 2007 Multilevel selection 2: Estimating the genetic parameters determining inheritance and response to selection. Genetics **175:** 289-299.

Bishop, S. C., and J. A. Woolliams, 2010 On the Genetic Interpretation of Disease Data. PLoS ONE **5:** e8940.

Chen, J., F. C. Michel, S. Sreevatsan, M. Morrison and Z. T. Yu, 2010 Occurrence and Persistence of Erythromycin Resistance Genes (erm) and Tetracycline Resistance Genes (tet) in Waste Treatment Systems on Swine Farms. Microbial Ecology **60:** 479-486.

Dawood, F. S., S. Jain, L. Finelli, M. W. Shaw, S. Lindstrom *et al.*, 2009 Emergence of a Novel Swine-Origin Influenza A (H1N1) Virus in Humans Novel Swine-Origin Influenza A (H1N1) Virus Investigation Team. New England Journal of Medicine **360:** 2605-2615.

Demeler, J., U. Kuttler, A. El-Abdellati, K. Stafford, A. Rydzik *et al.*, 2010 Standardization of the larval migration inhibition test for the detection of resistance to ivermectin in gastro intestinal nematodes of ruminants. Veterinary Parasitology **174:** 58-64.

Dempster, E. R., and I. M. Lerner, 1950 Heritability of threshold characters Genetics **35:** 212-236.

Denison, R. F., E. T. Kiers and S. A. West, 2003 Darwinian agriculture: When can humans find solutions beyond the reach of natural selection? Quarterly Review of Biology **78:** 145-168.

Doeschl-Wilson, A. B., R. Davidson, J. Conington, T. Roughsedge, M. R. Hutchings *et al.*, 2011 Implications of Host Genetic Variation on the Risk and Prevalence of Infectious Diseases Transmitted Through the Environment. Genetics **188:** 683-693.

Doeschl-Wilson, A. B., D. Vagenas, I. Kyriazakis and S. C. Bishop, 2008 Exploring the assumptions underlying genetic variation in host nematode resistance (Open Access Publication). Genetics Selection Evolution **40:** 241-264.

Gillespie, D. T., 1977 Exact stochastic simulation of coupled chemical-reactions. Journal of Physical Chemistry **81:** 2340-2361.

Gilmour, A. R., B. J. Gogel, B. R. Cullis and R. Thompson (Editors), 2006 *ASReml User Guide Release 2.0*. VSN International Ltd, Hemel Hempstead, UK.

Griffing, B., 1967 Selection in reference to biological groups. I. Individual and group selection applied to populations of unordered groups Australian Journal of Biological Sciences **20:** 127-&.

Griffing, B., 1976a Selection in reference to biological groups. 4. Use of extreme forms of nonrandom groups to increase selection efficiency Genetics **82:** 723-731.

Houston, R. D., C. S. Haley, A. Hamilton, D. R. Guy, J. C. Mota-Velasco *et al.*, 2010 The susceptibility of Atlantic salmon fry to freshwater infectious pancreatic necrosis is largely explained by a major QTL. Heredity **105:** 318-327.

Ihaka, R., and R. Gentleman, 1996 R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics **5:** 299-314.

Keeling, M. J., and P. Rohani, 2008 *Modelling Infectious Diseases in Humans and Animals*. Princeton University Press, Princeton

Keeling, M. J., and J. V. Ross, 2008 On methods for studying stochastic disease dynamics. Journal of the Royal Society Interface **5:** 171-181.

Lloyd-Smith, J. O., S. J. Schreiber, P. E. Kopp and W. M. Getz, 2005 Superspreading and the effect of individual variation on disease emergence. Nature **438:** 355-359.

Moore, A. J., E. D. Brodie and J. B. Wolf, 1997 Interacting phenotypes and the evolutionary process .1. Direct and indirect genetic effects of social interactions. Evolution **51:** 1352-1362.

Muir, W. M., 2005 Incorporation of competitive effects in forest tree or animal breeding programs. Genetics **170:** 1247-1259.

Muir, W. M., and J. V. Craig, 1998 Improving animal well-being through genetic selection. Poultry Science **77:** 1781-1788.

Nath, M., J. A. Woolliams and S. C. Bishop, 2004 Identifying critical parameters in the dynamics and control of microparasite infection using a stochastic epidemiological model. Journal of Animal Science **82:** 384-396.

Pullinger, G. D., S. C. Carnell, F. F. Sharaff, P. M. Van Diemen, F. Dziva *et al.*, 2010 Norepinephrine Augments Salmonella enterica-Induced Enteritis in a Manner Associated with Increased Net Replication but Independent of the Putative Adrenergic Sensor Kinases QseC and QseE. Infection and Immunity **78:** 372-380.

Raberg, L., D. Sim and A. F. Read, 2007 Disentangling genetic variation for resistance and tolerance to infectious diseases in animals. Science **318:** 812-814.

Ramirez-Valverde, R., I. Misztal and J. K. Bertrand, 2001 Comparison of threshold vs linear and animal vs sire models for predicting direct and maternal genetic effects on calving difficulty in beef cattle. Journal of Animal Science **79:** 333-338.

Van Dyken, J. D., T. A. Linksvayer and M. J. Wade, 2011 Kin Selection–Mutation Balance: A Model for the Origin, Maintenance, and Consequences of Social Cheating. The American Naturalist **177:** 288-300.

Vazquez, A. I., D. Gianola, D. Bates, K. A. Weigel and B. Heringstad, 2009 Assessment of Poisson, logit, and linear models for genetic analysis of clinical mastitis in Norwegian Red cows. Journal of Dairy Science **92:** 739-748.

Wilson, A. J., U. Gelin, M. C. Perron and D. Reale, 2009 Indirect genetic effects and the evolution of aggression in a vertebrate system. Proceedings of the Royal Society B-Biological Sciences **276:** 533-541.

Woolhouse, M. E. J., C. Dye, J. F. Etard, T. Smith, J. D. Charlwood *et al.*, 1997 Heterogeneities in the transmission of infectious agents: Implications for the

design of control programs. Proceedings of the National Academy of Sciences of the United States of America **94:** 338-342.

Yates, A., R. Antia and R. R. Regoes, 2006 How do pathogen evolution and host heterogeneity interact in disease emergence? Proceedings of the Royal Society B-Biological Sciences **273:** 3075-3083.

Chapter 3. Bias, accuracy and impact of indirect genetic effects in infectious diseases

The following chapter has been published in Frontiers in Livestock Genomics at the following URL:

http://www.frontiersin.org/Livestock_Genomics/10.3389/fgene.2012.00215/abstract

Recent advances in high throughput genomic information has led to new opportunities for dissecting genetic variation and to accelerate genetic improvement rendering control strategies for infectious diseases using selective breeding highly desirable. However, a major barrier to closing the genotype – phenotype gap is uncovering the genetic variance underlying disease phenotypes. To do so, as previously mentioned, genetic analyses require large sample sizes and hence disease phenotypes often need to be obtained from field data. Bishop and Woolliams (2010) have demonstrated that shortcomings of current estimation methods for field data which fail to take epidemiological considerations into account cause seemingly low heritability estimates for disease traits in domestic livestock. In the previous Chapter, it was demonstrated that conventional statistical models used for variance component estimation cannot capture genetic variation in host infectivity, when present in disease data, as they consider exposure as an environmental factor.

Host infectivity is the propensity of an infected individual to infect its group mates. The lack of attention to host variation in infectivity in genetic studies stands in stark contrast to the well-recognized important role of host infectivity in epidemiology. There is abundant evidence that heterogeneity in infectivity is ubiquitous, super-shedders being an extreme example**,** and can profoundly impact upon disease prevalence in the population (Doeschl-Wilson *et al.* 2011; Lloyd-Smith *et al.* 2005;

Woolhouse *et al.* 1997; Yates *et al.* 2006). However, it is not known to what extent infectivity is genetically controlled as it is difficult to measure directly. Evolutionary arguments would however suggest that there should be a significant amount of genetic variation in infectivity, because infectivity is not a component of an individual's fitness. Accumulation of genetic variation will, therefore, not be prevented by natural selection (Denison *et al.* 2003).

Disease data from field studies is often binary, indicating whether an individual became infected or not following exposure to infectious pathogens. Numerous studies have demonstrated that from this data one can infer genetic variation in individuals' underlying susceptibility to that disease (e.g. Houston *et al.* 2010). However, as demonstrated in the previous Chapter, standard models don't lend themselves to estimating genetic variation in infectivity as the effect of infectivity is observed in a different individual than the one expressing it. The theory of indirect genetic effects (IGE), also known as associative or social genetic effects, provides an appropriate framework to account for genetic variation in infectivity as it investigates heritable effects of an individual on the trait value of another individual (Griffing 1967). In this context, host infectivity can be regarded as an indirect effect to disease status.

The study presented in Chapter 2, was the first to demonstrate that genetic variation in host infectivity can be captured to some extent from binary disease data using an IGE model. However the results of that study suggest that there are severe shortcomings in using the standard IGE model, to estimate genetic variance in infectivity. For example the standard IGE model assumes that all individuals express the indirect effect (infectivity) at all times and that all individuals are affected by the indirect effect. However, only individuals who are infected can express infectivity. Furthermore, the infectivity of an infected individual will affect the disease status of susceptible group members only. Moreover, the number of both susceptible and infected individuals will change over time. In this way, the gross underestimation of genetic variance in infectivity observed in Chapter 2 may have occurred because the indirect genetic effect was attributed to all individuals in a group when in reality it

was expressed by and affecting only a subset of group members. Our hypothesis is, therefore, that an IGE model, when used to analyse binary disease data, may be improved by accounting for disease dynamics.

Here we explore the implementation of dynamic properties within the remit of a conventional quantitative genetics mixed model framework and software (ASReml, Gilmour 2006). To do so we specify which individuals contribute to an effect using the incidence matrix. In this study, two adjustments to the standard statistical IGE model were made. The first model, denoted the *Case* model, accounted for the fact that only infected individuals can express infectivity. The second model, denoted the *Case-ordered* model, also accounted for the fact that infected individuals can only affect individuals who didn't become infected before them. To evaluate these adjusted statistical models, we modelled disease progression in populations with genetic variation in host infectivity and susceptibility and estimated the genetic (co)variances in the simulated binary disease data with each model. The populations were simulated with varying epidemiological characteristics in order to assess their impact on our estimates. Finally, we evaluated the bias, accuracy and impact of the estimates obtained with each model and compared them to those obtained with the Standard (unadjusted) IGE model.

3.1 Materials & methods

3.1.1 The statistical models

3.1.1.1 Standard IGE model

The Standard IGE model has been described by Muir et al. (2005). Thus for disease phenotype *y* (e.g. infected or not) observed in individual *j* living in group *h* of size *n* with group mates *m*,

$$y_{jh} \sim \text{mean} + (\text{direct effect})_j + \sum_{m=1}^{n-1} (\text{indirect effect})_{mh} + (\text{group effect})_h + e_{jh}. \quad (3.1)$$

Bijma et al. (2007b) demonstrated the integration of this model into the quantitative genetics mixed models framework to obtain estimates for genetic variances and covariances of direct and indirect genetic effects. In the context of infectious disease, the direct effect relates to host susceptibility and the indirect effect to host infectivity. In the statistical analysis, the direct, indirect and group effects were all fitted as random effects. According to equation (3.1), for the Standard IGE model it is assumed that the direct and indirect effects are expressed by all individuals (i.e. expression does not depend on the disease status of an individual or that of its group mates).

3.1.1.2 Case IGE model

For the Case IGE model, model (3.1) was expanded to account for the fact that only infected individuals can express the indirect effect,

$$y_{jh} \sim \text{mean} + (\text{direct effect})_j + \sum_{m=1}^{n-1} X_{mh}(\text{indirect effect})_{mh} + (\text{group effect})_h + e_{jh}. \quad (3.2)$$

Where the indicator trait X_{mh} is equal to one if *m* became infected during the recording period and zero otherwise. In this way the number of individuals contributing to the IGE ($\sum_{m=1}^{n-1} X_{mh}$) will be equal to the number of group mates that have become infected during the observation period, representing the group's total exposure. The number of infected individuals will vary between groups not only for genetic reasons, but also due to environmental factors or chance. This creates a non-genetic covariance among group mates, which is accounted for by the random group effect. It is assumed that the population is naïve at the start of the recording period and that all individuals express susceptibility, although to varying extent (see section 3.1.2.2.).

3.1.1.3 Case-ordered IGE model

For the Case-ordered IGE model the Case model was expanded to include the order of infection of individuals, thus accounting for the fact that an infected individual *m* can only impact on group members that did not become infected prior to its own infection,

$$y_{jh} \sim \text{mean} + \psi \sum_{m=1}^{n-1} X_{m_j h} + (\text{direct effect})_j + \sum_{m=1}^{n-1} X_{m_j h}(\text{indirect effect})_{mh} + (\text{group effect})_h + e_{jh}. \quad (3.3)$$

The indicator trait $X_{m_j h}$ is equal to one if the group mate *m* became infected before individual *j*. The number of individuals contributing to the IGE ($\sum_{m=1}^{n-1} X_{m_j h}$), i.e. the exposure faced by individual *j*, will now vary between group mates and has n-1 possible levels. To account for differences in exposure between group mates, the effect of individual exposure ($\psi \sum_{m=1}^{n-1} X_{m_j h}$) was fitted as a separate fixed effect.

3.1.1.4 Variance structure

It was assumed that group mates are unrelated and all effects are independent of the residuals. The Standard IGE model can be written in the form of the Case IGE (3.2) with a constant indicator trait $X_{mh} = 1$ which has an expectation of one and zero variance. Assuming that all effects are independent of the residuals and given that $E(X^2) = E(X)$ as X is binary, the phenotypic variance for all three models can be partitioned, for a given level of individual exposure, as follows:

$$\sigma_y^2 = \sigma_d^2 + (n-1)E(X)\sigma_i^2 + \sigma_{group}^2 + \sigma_e^2. \quad (3.4)$$

Where *d* stands for the direct and *i* for the indirect effect. Thus the phenotypic variance for all three models differs only in the components pertaining to the

indicator trait *X*, i.e. *E(X)* which is the proportion of group mates expected to contribute to the IGE.

For the model fitting it was assumed that the vector of observed traits **y** follows a multi-variate normal distribution with means given by the fixed effects and the following variance structure:

$$\mathrm{Var}(\mathbf{y}) = \mathbf{Z}_a\mathbf{G}_a\mathbf{Z}_a' + \mathbf{Z}_e\mathbf{G}_e\mathbf{Z}_e' + \mathbf{R} \quad (3.5)$$

Where **R** and $\mathbf{G_e}$ are diagonal matrices with the residual and group variances, respectively, on the diagonal. $\mathbf{G_a}$ is the genetic (co)variance matrix and is given by the Kronecker product of a two by two variance-covariance matrix of direct and indirect effects and the relationship (**A**) matrix. $\mathbf{Z_a}$ is an incidence matrix linking individuals to their direct and indirect effects and $\mathbf{Z_e}$ an incidence matrix linking individuals to their group. Thus at each individual level, term four of equation 3.4 is given on the diagonal of **R** and term three on the diagonal of $\mathbf{G_e}$. As individuals are expected to be unrelated within groups, the direct-indirect covariance should not contribute to the phenotypic variance. The incidence matrix $\mathbf{Z_a}$ links one direct effects variance from $\mathbf{G_a}$ to the phenotypic variance, i.e. term one of equation 3.4, and $\sum_{m=1}^{n-1} X_{m_j h}$ indirect effects variances. Hence equation 3.4 is the expectation of the phenotypic variances given by equation 3.5.

3.1.2 Simulated Data

To evaluate the three models, simulated binary disease data were generated. For this purpose, epidemics were simulated with genetic variation in host susceptibility and infectivity following known distributions.

3.1.2.1 The epidemiological model

An epidemic was simulated in a population consisting of many groups (see section 3.1.2.2. below). The simulation describes disease progression in each group and

provides as output the disease status of each individual at given time points. These provided the binary disease records used for fitting the statistical models described in section 3.1.1. To avoid overburdening the results with unnecessary complexity we chose a simple compartmental stochastic susceptible-infected-recovered (SIR) model as detailed in Chapter 2. In an SIR model, individuals can be in one of three disease states, being susceptible (S), infected (I) or recovered (R). Individuals move through states in the order S → I → R. Initially, all individuals are in the S-state. Upon infection, a susceptible individual moves from the S-state to the I-state. Upon recovery, an infected individual moves to the R-state. The average rate of transition between the epidemiological compartments S, and I is determined by the transmission parameter β, whereas the average rate of transition between the compartments I and R is determined by the recovery rate γ. It was assumed that infected individuals become immediately infectious.

Genetic variation in host susceptibility and infectivity was incorporated into the model by assigning for each individual *j* its own level of susceptibility g_j and infectivity f_j. Hence, there is no longer a fixed transmission parameter β for the entire population, but the rate of transmission from individual *k* to individual *j* is given by the pair wise transmission parameter β_{jk}, which depends on the infectivity of *k* and the susceptibility of *j*. In order to reduce unnecessary noise it was assumed that variation in susceptibility and infectivity was fully genetic. However the outcome, i.e. whether an individual became infected or not, was assumed to be a stochastic event and will therefore contain both a genetic and a random non-genetic component. The same pair-wise transmission parameter β_{jk} was used as in Chapter 2 which was defined as

$$\beta_{jk} = -ln(1 - X_{g,j} g_j X_{f,k} f_k). \qquad (3.6)$$

Thus β_{jk} is a function of the product of the susceptibility *g* of individual *j* and the infectivity *f* of individual *k*. To reflect whether susceptibility is expressed by individual *j*, it is scaled by $X_{g,j}$ which equals one if *j* is susceptible and zero otherwise. Similarly, infectivity is scaled by $X_{f,k}$ which equals one if *k* is infected and

zero otherwise. In this way, transmission between individuals *j* and *k* can only occur if *j* is susceptible and *k* infectious. For simplicity no variation in individual speed of recovery γ_k was assumed. Hence, individual speeds of recovery γ_k were assumed to be equal to a constant γ if the individual was infected and zero otherwise.

The epidemic was simulated as a stochastic Poisson process as detailed in Chapter 2 which starts by infecting one randomly chosen individual in each group and describes disease progression in the population through a series of independent infection and recovery events. No transmission was assumed between groups.

3.1.2.2 Simulated populations

Similarly to Chapter 2, the simulated populations consisted of N=100,000 individuals with a paternal half-sib structure and no full sibs. All parents were assumed to be unrelated. The half sib family size was 100 individuals. Each population was divided into 10,000 groups of size n=10 chosen at random without reference to pedigree. Since each population was divided into 10,000 groups giving rise to 10,000 independent epidemics, each simulation was replicated only ten times. The simulation was run for populations with variation introduced in both susceptibility and infectivity. Note that none of the IGE models presented in this chapter take individuals' recovery into account. Therefore, to assess the impact of recovery speed, on the outcome of the subsequent analyses, the simulations were run for populations with constant speed of recovery $\gamma = 0.1, 0.01 \text{ or } 0.001$. These populations will be referred to as having a high, medium or low recovery rate, respectively.

Similarly to Chapter 2, breeding values for susceptibility and infectivity were assumed to be distributed according to the right-skewed gamma distribution $\Gamma(a,\theta)$. This distribution was chosen because the distribution of infectivity is often right-skewed (Lloyd-Smith *et al.* 2005). Moreover, skewed distributions allow for larger variation when the distribution is confined to positive values. The same parameters were chosen for both susceptibility and infectivity in order to allow for direct comparisons. Specifically, the parameters were taken as a=1.14 and θ=0.18, such that the mean $a\theta$=0.21 the variance $a\theta^2$=0.037 and the distribution is right-skewed

with skewness $2/\sqrt{a}$=1.87. For details on how the breeding values of the offspring generation were constructed from the parental generation such as e.g. Mendelian inheritance, please refer to Chapter 2.

A recent study showed molecular evidence for a positive correlation between susceptibility and infectivity as the known immunosuppressant stress hormone norepinephrine was shown to cause increase shedding of Salmonella (Pullinger *et al.* 2010). In order to examine the impact of such covariation between susceptibility and infectivity, the correlation between both parameters were either set to zero or 0.35. If no correlation was assumed, the breeding values were assigned to individuals as detailed in Chapter 2. Non-zero correlations were generated by assigning parental breeding values using the gamma trivariate reduction algorithm as specified in Schmeiser and Lal (1982).

3.1.3 Validation of the statistical models

3.1.3.1 Estimating genetic parameters from simulated data

Genetic parameters and breeding values associated with host susceptibility (direct effect) and infectivity (indirect effect) were estimated with the three statistical models presented in section 3.1.1. The phenotypes used for this purpose were binary records describing the disease state of the simulated individuals during a given recording time. The latter was chosen such that the mean number of infected individuals per group was approximately n/2 in all populations. The binary disease trait, denoted here as ‘disease presence’, was one if an individual had become infected during the recording time and zero otherwise.

The Case-Ordered model (3.3) not only required information on the disease state of individuals but also on the order of infection within each group. Infection occurs over a continuous time scale. However, in practice data is often recorded at discrete sampling times. Knowledge of the exact order would in practice be equivalent to dividing the recording period into an infinite number of sampling times. Here the Case-ordered model was simplified so that the recording time was split into two

sampling times. Thus the records used in the analyses were the disease presence of individuals recorded at the two sampling times. The length of each period was taken such that approximately half of the individuals that had become infected by the end of the recording time, would have become infected during the first period and the other half in the second period. The reasoning behind this choice in sampling times is outlined in the discussion. Thus the indicator trait X_{m_jh} for individual *j* with group mate *m* in equation (3.3) was equal to one only if group mate *m* had become infected and *j* was still susceptible at the start of the recording period in which *m* became infected.

The phenotypes of randomly chosen individuals initiating the epidemic in each group were removed prior to analysis. The genetic information of these individuals was however included in the analysis. Due to difficulties with convergence (see discussion) the direct-indirect covariance was fixed prior to analysis to zero when susceptibility and infectivity were independent and to 0.014, in correspondence with the simulated correlation, otherwise. All genetic analyses were carried out using ASReml (Gilmour 2006).

3.1.3.2 Validation criteria

3.1.3.2.1 Expected variance

In the simulations, genetic variation was introduced in the underlying parameters infectivity and susceptibility. The analysis of simulated data, however, was performed at the level of observed disease status (0, 1). To judge the results of the data analysis, i.e. to compare the estimated values to their expected values, it is necessary to transform susceptibility and infectivity to the observed binary scale. Following Dempster and Lerner (1950) we assumed a linear relationship between the susceptibility and the observed binary phenotype of an individual and between the infectivity of an individual and the binary phenotypes of its group mates. Specifically, the effect of susceptibility was obtained by regressing the individuals' phenotypes *y* on their susceptibility *g* ($y \sim b_1 g$). To obtain the relationship between infectivity and the disease status of an individual's group mates, the phenotypes *y* of

individuals *j* were regressed on the infectivity *f* of a randomly chosen infected group mate *k* of *j* ($y_{jh} \sim b_2 f_{kh}$). This approximation was used as there are many groups and relatively few group mates. The corresponding regression coefficients b_1 and b_2 were estimated using the statistical package R (Ihaka and Gentleman 1996) using the known breeding values and phenotypes from the simulations. Similarly to the genetic analysis, the phenotypes of the individuals which were randomly chosen to initiate the epidemic were discarded. In this way, the true direct breeding value of individual *j* (BVd_j) corresponds to $b_1 g_j$ and its indirect breeding value (BVi_j) to $b_2 f_j$. The expected (co)variances for the direct and indirect effects are then given by the (co)variances of their corresponding true BVs. The expected phenotypic variance was estimated as $\bar{p}(1-\bar{p})$, $\bar{p}$ denoting mean prevalence.

3.1.3.2.2 Bias and accuracy

Estimates of bias for both the direct and indirect effects were obtained by regressing the true BVs for direct and indirect effects (as derived above) on the corresponding estimated breeding values (EBVs) obtained from each model.

Accuracy was estimated as the correlations between the true breeding values (BVs) for susceptibility and infectivity with the corresponding direct and indirect effect EBVs. Note that transformation of susceptibility / infectivity BVs to binary scale BVs was not necessary for calculating correlations under the assumption of a linear relationship between the underlying and observed scales.

3.1.3.2.3 Impact of selection

In order to estimate the impact of the three models on response to selection the population true mean susceptibility/infectivity was compared to the true mean susceptibility/infectivity after selection of 10% of the individuals with the lowest EBVs obtained from each model. For all three models, selection was carried out

based upon the EBVs for direct effect (EBVd) and indirect genetic effect (EBVi) separately as well as for the index $I_x = \mathrm{EBVd} + x\mathrm{EBVi}$ with *x* taken as the product of the expected number of individuals contributing to the IGE and the expected total group exposure rounded to the nearest integer. Specifically, *x*=4, 2 and 1 for the Standard, Case and Case-ordered model respectively. These weights were chosen to take into account the number of individuals contributing to the indirect effect as well as the level of exposure. Moreover, they provided the greatest impact when tested on the population with zero correlation between susceptibility and infectivity and recovery rate 0.01.

To quantify response to selection in terms of risk and severity of the epidemic, the basic reproduction number R_0 was estimated for the whole population and for each selected subpopulation using the true values of susceptibility and infectivity from the simulation. R_0 is the mean number of secondary infections an infected individual will cause in its lifetime in an otherwise naïve population, and is commonly used as a measure of disease risk and severity in epidemiology (Anderson and May 2006). By definition an epidemic will die out if $R_0<1$. Following a SIR model for a closed population, $R_0 = \beta S_0/\gamma,\ S_0 = (n-1)$ being the initial number of susceptible individuals in a group (Keeling and Rohani 2008). Incorporating equation (3.6) and taking a Taylor series expansion we obtain as zero order approximation:

$$R_0 \approx \left(-\ln(1-\bar{g}\bar{f})S_0\right)/\gamma. \quad (3.7)$$

3.2 Results

All results presented in sections 3.2.1-3.2.3 refer to the populations with a zero correlation between infectivity and susceptibility.

3.2.1 Variance estimates

Table 3-1 shows the variance estimates obtained for each population, from all three models, along with the expected variances. Overall the variance estimates obtained with the Case model are in best agreement with the expected variances. This is particularly true for the populations with a medium to slow recovery rate. Whilst the Standard model vastly underestimates the indirect genetic effects variance, the Case-ordered model provides vastly inflated estimates. Moreover, in contrast to the Standard and Case models, the Case-ordered model also grossly underestimates the direct effects variance. From this it is clear that the Case-ordered model is inadequate. Therefore further results for the Case-ordered model will not be shown as they merely confirmed the overall poor performance of this model. Potential explanations and alternative suggestions are outlined in the discussion section of this chapter.

The indirect effects variance estimate obtained with the Case model deviates most from the expected variance in the population with a high recovery rate (γ=0.1) i.e. when infected individuals are most likely to recover during the epidemic. This is not surprising, as the Case model does not take the time period of infection into account. Thus, individuals that recover early would be assumed to contribute infectivity during the entire recording period. It is also noteworthy that the direct effects variance (both expected and estimated) increases and the indirect variance decreases with decreasing recovery rate. This demonstrates that the relative contributions of both effects to the overall variance strongly depend on epidemiological characteristics.

Table 3-1 Genetic variance estimates

Expected and estimated genetic variance in the direct and indirect effect in populations with a high, medium and low recovery rate. Variance components were estimated with the Standard, Case and Case-ordered models. All variance components have been scaled by 10^3.

Effect	Recovery rate γ	Model			
		Expected	Standard	Case	Case-ordered
Direct	0.1	20.61	20.28	23.84	9.65
	0.01	33.79	35.07	37.28	11.45
	0.001	34.30	35.26	36.73	9.03
Indirect	0.1	16.70	0.33	6.72	38.80
	0.01	8.25	0.72	6.32	86.47
	0.001	5.52	0.70	7.08	108.45

3.2.2 Bias and accuracy

Figure 3-1 shows the standardised bias estimates obtained for each population from the Standard and Case models. These results confirm the conclusions from the comparison between estimated and expected variance components (see Table 3-1). Specifically, the BV estimates obtained for the direct effect show little bias with either model. However, the Standard model grossly underestimates the indirect effect BV. It is noteworthy that, whilst the estimates obtained with the Case model show less bias overall, both the degree and direction of the standardised bias estimates depended on the recovery rate. Specifically, the standardised bias estimates for the indirect effect obtained with either model show an upward trend as the recovery decreases. This is in line with the results of section 3.2.1. showing that the expected variance in the indirect effect decreases with the recovery rate whereas the estimated variance in the indirect effect obtained with either the Standard or Case model remain more or less constant. This suggests that epidemiological characteristics affect the bias of indirect effect estimates and further improvements may be possible if these are properly accounted for.

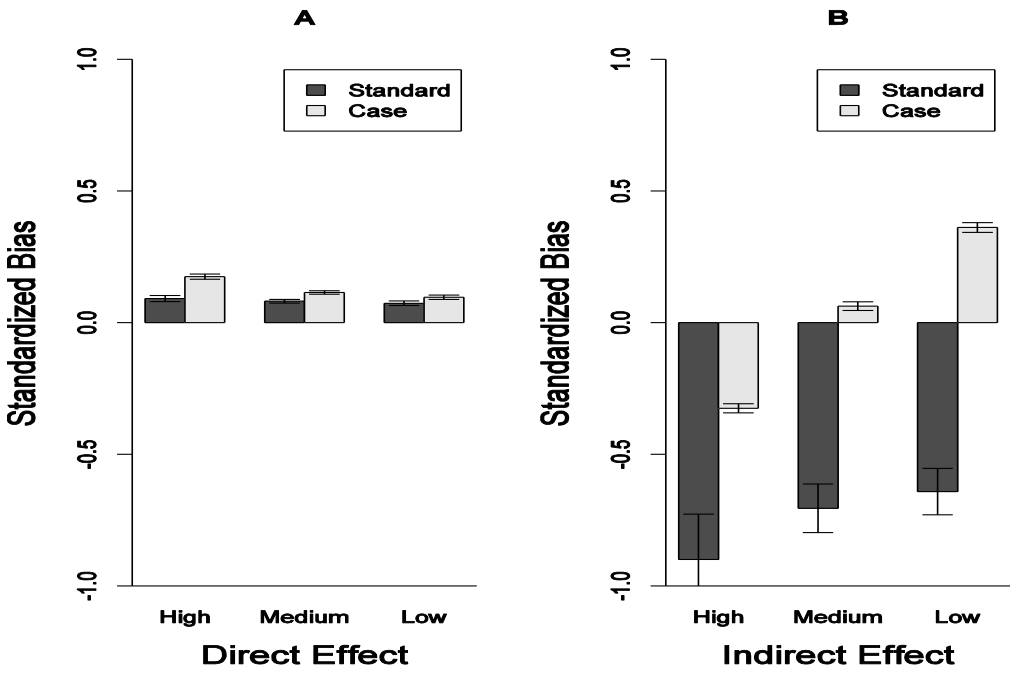

Figure 3-1 Bias of direct and indirect effect BV estimates for populations with different recovery rates (High, Medium, Low)

The bias estimates (regression coefficient of the true BVs on the EBVs), obtained for the Case and Standard model, were standardised to 1-bias if bias<1 and 1/bias-1 if bias>1, in order to show over and under estimation of the effects at the same scale. Thus values > 0 show over-estimation and values < 0 underestimation of the breeding values.

Figure 3-2 shows the accuracy estimates obtained for each population, for the Standard and Case models. The accuracy of the direct effect BV obtained for the Case model is similar to that obtained for the Standard model in all populations. However, the indirect effect BV estimates obtained with the Case model have a greater accuracy compared with those obtained with the Standard model. Also, there is a slight increase in the accuracy of the direct effect BV estimates obtained with the Standard and the Case model as the recovery rate decreases. This coincides with the increase in expected variance of the direct effect. It may also be due to the fact that for both the Standard and the Case model it is assumed that individuals express infectivity throughout the observation period. This assumption becomes more valid as individuals become less likely to recover, i.e. as the recovery rate decreases.

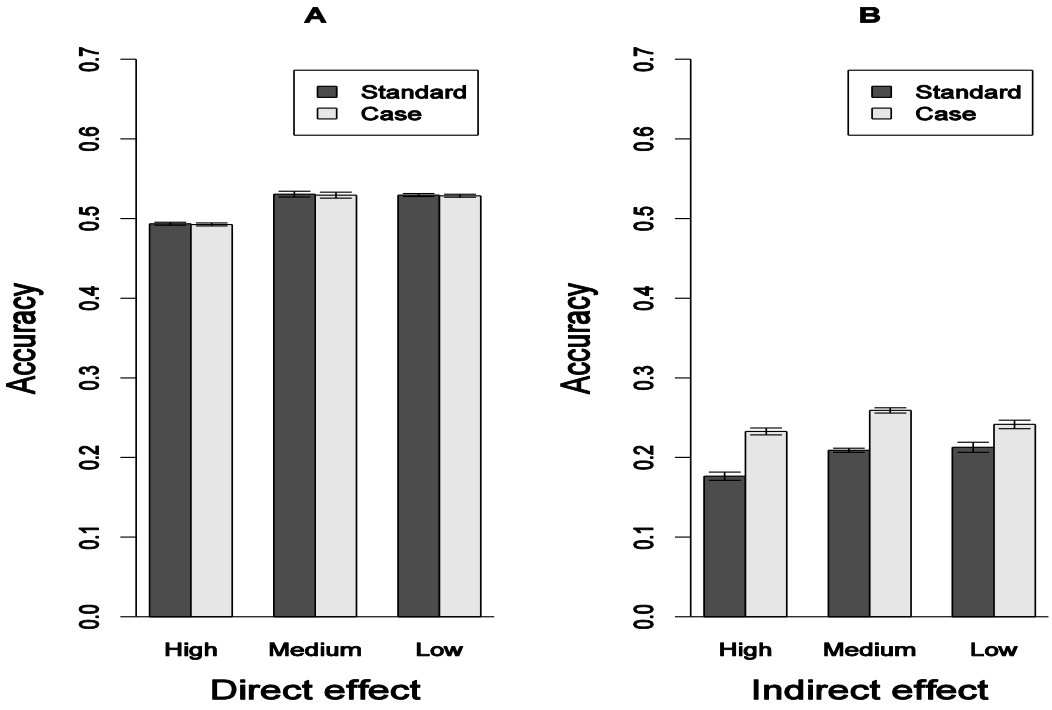

Figure 3-2 Accuracy of direct and indirect effect BV estimates for populations with different recovery rates (High, Medium, Low)

Note that the accuracy estimates obtained for the direct effect BVs with the Standard and Case model are reasonable given the half-sib structure of the population. The accuracy estimates obtained for the indirect effect BVs, on the other hand, are much lower, thus indicating that there is still further scope for improvement.

3.2.3 Impact of selection

Table 3-2 shows the true mean susceptibility and infectivity values and the basic reproduction number R_0 (scaled by γ) after selection using the EBVs obtained with the Standard and Case model from each population. Overall, selection using an index of direct and indirect EBVs obtained with the Case model shows slightly more reduction in risk and severity of an epidemic as measured by R_0. However, the

difference between the use of an index and the direct effect EBVs alone is small. This makes sense given the low accuracy estimates obtained for the indirect effect EBVs. The benefits of the Case model over the Standard model are mainly caused by the improved estimates for the indirect effects EBVs. Whilst selection on the direct EBVs from the Case model made little to no difference on true mean susceptibility compared with selecting on the direct EBVs from the Standard model, selection on the indirect EBVs from the Case model led to greater reduction of true mean infectivity.

It is noteworthy that the mean susceptibility increases, when selecting on the indirect effect EBVs from all analyses except for the Case model with a high recovery rate. This general increase in mean susceptibility can be explained by the fact that only infected individuals can express infectivity. Thus individuals with a low susceptibility are less likely to express infectivity. It is therefore less likely that the EBV for infectivity of these individual would be on the extreme (selected) ends of the distribution. However, there is a slight decrease, rather than increase, in mean susceptibility, when selecting on the indirect effect EBVs obtained with the Case model from the population with a high recovery rate. This may be explained by the following. As seen in section 3.2.1., the effect of infectivity increases as the recovery rate increases probably due to an increase in the importance of being infected early (high susceptibility). Hence, individuals with a high susceptibility are more likely to be assigned a high infectivity EBV as the recovery rate increases. This is in line with the fact that the mean susceptibility, when selecting on indirect effect EBVs obtained with the Case model, increases as the recovery rate decreases.

Table 3-2 Selection impact on true susceptibility, infectivity and risk and severity of an epidemic

For all three models, selection was carried out based upon the EBVs for direct effect (EBVd) and indirect genetic effect (EBVi) separately as well as for the index Ix = EBVd + xEBVi, with *x* taken as the product of the expected number of individuals contributing to the IGE and the expected total group exposure rounded to the nearest integer. Specifically, *x*=4 and 2 for the Standard and Case model respectively. Standard errors all $<5*10^{-3}$ unless indicated.

Recovery rate γ	Model	Selected on:	Mean susceptibility	Mean infectivity	$R_0*\gamma$
		No selection	0.21	0.21	0.41
0.1	Standard	EBV_d	0.07	0.22	0.15
		EBVi	0.24	0.15	0.34
		EBV_x	0.08	0.21	0.15
	Case	EBV_d	0.08	0.22	0.15
		EBVi	0.18	0.13	0.22
		EBV_x	0.08	0.19	0.14
0.01	Standard	EBV_d	0.08	0.21	0.15
		EBVi	0.24	0.15	0.33
		EBV_x	0.08	0.20	0.15
	Case	EBV_d	0.08	0.21	0.15
		EBVi	0.22	0.13	0.26
		EBV_x	0.08	0.19	0.14
0.001	Standard	EBV_d	0.08	0.21	0.15
		EBVi	0.23	0.15	0.32±0.01
		EBV_x	0.09	0.19	0.15
	Case	EBV_d	0.08	0.21	0.15
		EBVi	0.24	0.14	0.30
		EBV_x	0.08	0.19	0.14

3.2.4 Effect of dependence between susceptibility and infectivity

Having a positive correlation between susceptibility and infectivity had little to no impact on the bias of all estimates (results not shown) and the accuracy of the direct effect variance estimates. The accuracy of all indirect effect estimates, however, increased when there was a positive correlation between susceptibility and infectivity (Figure 3-3). This increase in the accuracy of the indirect effect EBVs may be due to

the fact that the accuracy of EBVs obtained by Best Linear Unbiased Prediction inherently improves when they are positively correlated (Falconer and Mackay 1996). Moreover, it may also be due to the fact that whether infectivity is expressed or not depends on susceptibility, even if infectivity and susceptibility themselves are independent. In that way, the indirect effect EBVs will partly depend on susceptibility and hence they will be more accurate if there truly is a positive correlation between infectivity and susceptibility.

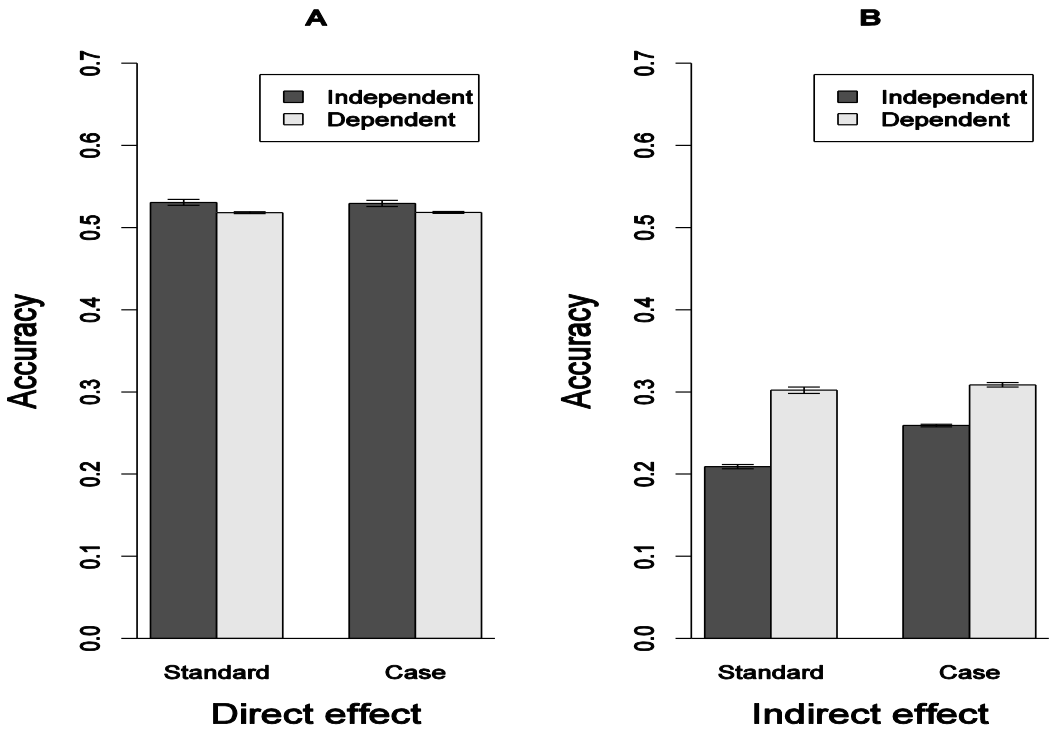

Figure 3-3 Accuracy of direct and indirect effect estimates in populations with(out) dependence between susceptibility and infectivity

Results shown for populations with a medium recovery rate, similar results were obtained for populations with different recovery rates. The correlation between susceptibility and infectivity is 0 in the independent population and 0.35 in the dependent population.

Finally, the impact of selection on true mean susceptibility, infectivity and R_0 is compared between populations with and without dependency between susceptibility

and infectivity, and a medium recovery rate, in Table 3-3. As may be expected, the greatest impact of selection on R_0 was obtained in the population with a positive correlation between susceptibility and infectivity. A similar improvement was observed in populations with other recovery rates (results not shown).

Overall, the performance of the Standard model was closer to that of the Case model when there was a positive correlation between susceptibility and infectivity. Note that in order to achieve convergence the covariance estimate was fixed in all analyses. Varying the value at which the covariance is fixed slightly affected the bias estimates but not the accuracy nor any of the previous observations.

Table 3-3 Selection impact in a population with a positive correlation between susceptibility and infectivity

For populations with a medium recovery rate γ=0.01. Standard errors all $<5*10^{-3}$.

		Mean susceptibility		Mean infectivity		R_0*γ	
	Correlation:	0	0.35	0	0.35	0	0.35
	No selection	0.21	0.21	0.21	0.21	0.41	0.41
Standard	EBV_d	0.08	0.08	0.21	0.14	0.15	0.10
	EBVi	0.24	0.10	0.15	0.13	0.33	0.12
	EBV_x	0.08	0.08	0.20	0.13	0.15	0.10
Case	EBV_d	0.08	0.08	0.21	0.14	0.15	0.10
	EBVi	0.22	0.15	0.13	0.13	0.26	0.17
	EBV_x	0.08	0.08	0.19	0.12	0.14	0.09

3.3 Discussion

Chapter 2 showed that IGE models developed for production traits provide a promising tool for estimating genetic variation underlying binary disease data. However, standard IGE models did not fully capture genetic variation in infectivity. The hypothesis of this study was that extending an IGE model to allow for disease dynamics ought to improve its ability to estimate genetic variation in susceptibility and infectivity from binary disease data. Here we explored the extent to which it is

possible to do so, within the remit of the conventional mixed model framework and software. In these conditions it is possible to specify the individuals contributing to the indirect effect using the incidence matrix. The effect of including disease dynamics, in this way, was assessed by comparing the accuracy, bias and selection impact of two adjusted IGE models, with the Standard IGE model, using simulated data. In the first adjusted IGE model, the Case model, it was assumed that only infected individuals have an indirect effect on their group mates. In the second adjusted IGE model, the Case-ordered model, it was assumed that infected individuals only have an indirect effect on susceptible group mates. Our results show that taking the disease status of individuals into account, by using the Case model, considerably improved the bias, and showed some improvement in accuracy and impact of genetic infectivity estimates from binary disease data compared to the Standard model. However, although heuristically one would assume that the Case-ordered model would provide the best estimates, as it takes most of the disease dynamics into account, in fact it provides the worst.

The poor performance of the Case-ordered model reveals that the straight-forward approach of incorporating information about disease dynamics in the form of incidence matrices into a linear mixed model has severe limitations. The problem with using incidence matrices containing information about individuals contributing to the indirect effect as explanatory variables (i.e. on the right-hand-side of the statistical model) is that they use information obtained from the very observations they try to predict. Thus, the indicator traits, like the observations, are partly determined by the breeding values we are trying to estimate. Note that this was also true for the Case model, but to a much lesser extent as the indicator trait corresponds to the disease state of another individual in that model. However, the indicator trait in the Case-ordered model is a property of the susceptibility of two individuals, potentially rendering the numerical relationships too complex for the estimation software used. In fact, a much simplified approximation had to be used for the Case-ordered model in order to achieve convergence. Indeed, the number of observation periods had to be reduced to two. Finally, there may be a reduction in statistical power as the number of individuals contributing to the indirect effect decreases. Thus

a different approach is needed for implementing information about the order of infection into the statistical models. For example, hierarchical Bayesian models may provide a better framework for incorporating infection order in terms of prior information.

Our results reveal that the contribution of susceptibility and infectivity to an individual's disease status, as well as the bias and accuracies of the corresponding EBVs obtained with either model, depend on epidemiological characteristics. In particular, the expected direct effect variance will be more important in diseases with a low recovery rate and the expected indirect effect variance will be more important in diseases with a high recovery rate. This may be due to the fact that, when recovery is slow, the exposure will be relatively high as individuals remain infected for longer. Thus not getting infected is more likely the result of a low susceptibility, increasing the relative contribution of variance in susceptibility to the total phenotypic variance. When recovery is fast, on the other hand, having a sufficiently high infectivity in order to spread the infection prior to recovering becomes more important. In accordance with the greater relative contribution of variance in susceptibility, the accuracies of the direct effects EBV are also slightly higher when recovery rate is low. However, the accuracies of the indirect EBVs obtained with the Standard and Case models are decreased in the population with a fast recovery compared with those with a medium or slow recovery. This may be due to the fact that both these models assume a constant expression of infectivity. Hence the assumptions underlying these models are more accurate in populations with a slow recovery. It must also be noted that, including the individuals initiating the epidemic, only about 50% of individuals became infected in all populations. This should be good for estimating the variance in the direct effect but it does mean that approximately 50% of individuals never express infectivity. Further work is therefore required to evaluate the optimal recording time given epidemiological parameters such as recovery rate.

Our results indicate that a positive correlation between susceptibility and infectivity improves the EBVs obtained with all three models in terms of accuracy and impact

of selection. The gain in selection impact, when selecting on an index of direct and indirect EBVs, is somewhat expected as selection on either EBV will also be expected to affect response in the other due to dependency. Moreover, it has repeatedly been demonstrated that the covariance between direct and indirect effect is a component of the expected response to selection when using an IGE model (Griffing 1967). The gain in accuracy of the indirect effect EBVs probably occurs because an individual must be infected, which depends on that individual's susceptibility, in order to express infectivity. However, these results stem from a specific correlation value and it may be worth investigating whether different correlation values would affect this trend. Similarly the effect of different experimental settings such as grouping related vs non-related individuals would be worth investigating as they have been demonstrated to strongly affect the scale of parameter estimates (Bijma 2010).

It must be noted that the model validation partly depended on expected variances and BVs on a binary scale. In this study a simple linear relationship was assumed, following Dempster and Lerner (Bijma *et al.* 2007; 1950), between the observed binary trait and the underlying genetic parameters. Alternatively, we could have linked the linear mixed model describing the underlying parameters to the binary trait with a non-linear link function using a generalized linear mixed model GLMM. However, the relationship between the underlying genetic parameters and the observed disease status is complex and stochastic. It is therefore unlikely that canonical link functions, such as the probit or logit function, are appropriate in our case. In fact, in Chapter 2 we demonstrated that using a GLMM linking the Standard IGE model with our binary disease trait with the logit function provided qualitatively similar results to those obtained without the transformation. Moreover, there was no advantage in using such a transformation as the relationship was not only inappropriate, but it also provided intractable estimates and seemed to increase the interaction bias. In the following Chapter we establish the appropriate relationship between the underlying genetic parameters and the observed binary host infectious disease data. By doing so we demonstrate that the probit and logit link functions are inappropriate for the analysis of binary host infectious disease data. Moreover, we

demonstrate that the use of a complementary log-log link function or survival analysis is useful when there is variation in susceptibility only but inadequate if there is variation in infectivity. The relationship established in Chapter 4 cannot be readily implemented into existing software and is therefore beyond the scope of this study investigating the incorporation of disease dynamics within the framework of a conventional quantitative genetics mixed model and associated software. We therefore decided to use a linear mixed model, which have been shown to provide estimates of genetic parameters of sufficient accuracy to generate selection response (e.g. Vazquez *et al.* 2009). Other types of models which may be interesting to adapt and develop further in order to estimate genetic parameters associated with host susceptibility and infectivity include cure models (Ødegård *et al.* 2011) and product threshold model (David *et al.* 2009). Cure models, have the potential to consider expression of infectivity conditional on infection status, whereas product threshold models might better represent the interaction between a susceptible and infectious individual.

In summary, we have shown that epidemiological characteristics and disease dynamics strongly influence estimates of genetic variances and breeding values associated with host susceptibility and infectivity and thus cannot be ignored. The straight-forward approach of incorporating dynamic information in the form of incidence matrices into the mixed model framework using conventional animal breeding software is appealing due to ease of use and general availability and showed improvement over the standard IGE model. However this approach also has substantial limitations in incorporating disease dynamics. An alternative approach for incorporating epidemiological information and dynamic aspects would entail establishing an appropriate mathematical function that links the binary disease trait to underlying epidemiological parameters under genetic influence and encapsulates dynamic aspects.

3.4 References

Anderson, R. M., and R. M. May, 2006 *Infectious Diseases of Humans*. Oxford University Press, Oxford.

Bijma, P., 2010 Estimating Indirect Genetic Effects: Precision of Estimates and Optimum Designs. Genetics **186:** 1013-1028.

Bijma, P., W. A. Muir and J. a. M. Van Arendonk, 2007 Multilevel selection 1: Quantitative genetics of inheritance and response to selection. Genetics **175:** 277-288.

Bishop, S. C., and J. A. Woolliams, 2010 On the Genetic Interpretation of Disease Data. Plos One **5**.

David, I., L. Bodin, D. Gianola, A. Legarra, E. Manfredi *et al.*, 2009 Product versus additive threshold models for analysis of reproduction outcomes in animal genetics. Journal of Animal Science **87:** 2510-2518.

Dempster, E. R., and I. M. Lerner, 1950 Heritability of threshold characters Genetics **35:** 212-236.

Denison, R. F., E. T. Kiers and S. A. West, 2003 Darwinian agriculture: When can humans find solutions beyond the reach of natural selection? Quarterly Review of Biology **78:** 145-168.

Doeschl-Wilson, A. B., R. Davidson, J. Conington, T. Roughsedge, M. R. Hutchings *et al.*, 2011 Implications of Host Genetic Variation on the Risk and Prevalence of Infectious Diseases Transmitted Through the Environment. Genetics **188:** 683-693.

Falconer, D. S., and T. F. C. Mackay (Editors), 1996 *Introduction to Quantitative Genetics*. Pearson Education Limited, Harlow.

Gilmour, A. R. G., B.J.; Cullis, B.R.; Thompson, R., 2006 *ASReml User Guide Release 2.0*. VSN International Ltd, Hemel Hempstead.

Griffing, B., 1967 Selection in reference to biological groups .I. Individual and group selection applied to populations of unordered groups. Australian Journal of Biological Sciences **20:** 127-&.

Houston, R. D., C. S. Haley, A. Hamilton, D. R. Guy, J. C. Mota-Velasco *et al.*, 2010 The susceptibility of Atlantic salmon fry to freshwater infectious pancreatic necrosis is largely explained by a major QTL. Heredity **105:** 318-327.

Ihaka, R., and R. Gentleman, 1996 R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics **5:** 299-314.

Keeling, M. J., and P. Rohani, 2008 *Modelling Infectious Diseases in Humans and Animals*. Princeton University Press, Princeton

Lloyd-Smith, J. O., S. J. Schreiber, P. E. Kopp and W. M. Getz, 2005 Superspreading and the effect of individual variation on disease emergence. Nature **438:** 355-359.

Muir, W. M., 2005 Incorporation of competitive effects in forest tree or animal breeding programs. Genetics **170:** 1247-1259.

Ødegård, J., P. Madsen, R. Labouriau, B. Gjerde and T. H. E. Meuwissen, 2011 A sequential threshold cure model for genetic analysis of time-to-event data. Journal of Animal Science **89:** 943-950.

Pullinger, G. D., S. C. Carnell, F. F. Sharaff, P. M. Van Diemen, F. Dziva *et al.*, 2010 Norepinephrine Augments Salmonella enterica-Induced Enteritis in a Manner Associated with Increased Net Replication but Independent of the

Putative Adrenergic Sensor Kinases QseC and QseE. Infection and Immunity **78:** 372-380.

Schmeiser, B. W., and R. Lal, 1982 BIVARIATE GAMMA-RANDOM VECTORS. Operations Research **30:** 355-374.

Vazquez, A. I., D. Gianola, D. Bates, K. A. Weigel and B. Heringstad, 2009 Assessment of Poisson, logit, and linear models for genetic analysis of clinical mastitis in Norwegian Red cows. Journal of Dairy Science **92:** 739-748.

Woolhouse, M. E. J., C. Dye, J. F. Etard, T. Smith, J. D. Charlwood *et al.*, 1997 Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. Proceedings of the National Academy of Sciences of the United States of America **94:** 338-342.

Yates, A., R. Antia and R. R. Regoes, 2006 How do pathogen evolution and host heterogeneity interact in disease emergence? Proceedings of the Royal Society B-Biological Sciences **273:** 3075-3083.

Chapter 4. A unifying theory for genetic epidemiological analysis of binary disease data

The following chapter has been published in GSE at the following URL:

http://www.gsejournal.org/content/46/1/15

With the rapid accumulation of data on the genetic regulation of host responses to infectious pathogens, the drive towards strategies that control genetic disease is gaining momentum. Genetic approaches to combat infectious disease tend to focus on improving host resistance, i.e. the ability of a host to block pathogen entry or to counteract pathogen replication within the host. However, despite enormous breakthroughs in genomics, estimating genetic parameters for disease resistance has proven considerably more challenging than analysis of production traits, and this has hampered the incorporation of disease traits into breeding programmes. These challenges partly arise because disease resistance is not a trait that is directly measurable but relies on observable proxies. Due to the requirement of large sample sizes for quantitative genetic analyses, such proxies are often obtained from field data, which are typically binary, indicating whether an individual has become infected or not (Bishop and Woolliams 2010).

Current quantitative genetic methods analyse binary infectious disease data essentially by contrasting the set of individuals diagnosed as infected to those diagnosed as non-infected, assuming that the observed phenotypic differences represent differences in host resistance to the pathogens under consideration (Falconer and Mackay 1996). However, the corresponding statistical models, such as threshold or logit models, entail several intrinsic assumptions that are unrealistic in the case of infectious disease. First, the observations (e.g. diseased / not diseased) are

assumed to be accurate but in reality, the diagnostic tools that are used in the field rarely have complete sensitivity or specificity, i.e. there is a considerable chance for misclassification of individuals as healthy or diseased. Second, it is assumed that exposure to infectious pathogens of individuals that share the same environment is (a) equal between individuals, (b) constant over time and (c) purely environmental. However, in large groups with a non-uniform contact structure, there may be substantial heterogeneity in exposure at any given time. Thus, an individual classed as healthy may have indeed greater resistance, or could simply be misdiagnosed, or may not yet have come in contact with the infectious agents. Furthermore, for infectious diseases transmitted by direct contact, the disease status of an individual is not just the expression of its own resistance in a constant infectious environment. Instead infections result from dynamic interactions between susceptible and infected individuals, and genetic variation may be inherent to all such interactions. As the number of infected individuals in a population changes throughout the time course of a disease outbreak, exposure will change as well. Lastly, exposure depends on how infectious the infected individuals are, which may differ between individuals, e.g. due to different shedding patterns of infectious material or different durations of shedding. Thus, not only host resistance but also host infectiousness, i.e. the ability of a host to transmit an infection, may display substantial host genetic variation.

All of the above characteristics that are inherent to natural disease outbreaks are likely to affect estimates of genetic parameters for disease traits. Indeed, in chapters 2 & 3 we demonstrated that conventional quantitative genetics models fail to capture host genetic variation in infectiousness, if present. Furthermore, theoretical work has established that imperfect diagnostics and incomplete or variable exposure produce a downward bias in estimates of heritability and of SNP (single nucleotide polymorphism) effects, and affect inferences about modes of inheritance of SNP effects for disease resistance (Bishop *et al.* 2012; Bishop and Woolliams 2010). This theory is empirically supported by comparing results from recent field and challenge experiments that aimed at estimating genetic parameters and at identifying genetic markers for the resistance of pigs to the Porcine Reproductive and Respiratory Syndrome Virus (PRRSV) (Biffani *et al.* 2010; Boddicker *et al.* 2012). Both these

studies included approximately 1200 animals, but whereas infection resulted from natural transmission dynamics in the field studies (Biffani *et al.* 2010), the challenge experiment infected all animals with the same dose of a particular PRRSV strain (Boddicker *et al.* 2012), thus excluding the various sources of heterogeneity in exposure outlined above. In accordance with theory, heritability estimates for viraemia were considerably lower based on field data than from challenge data (0.096 vs. 0.31) and the challenge study found a major QTL for disease resistance that had not been identified in the field data. Thus, both theory and experimental evidence imply that, in order to use data from natural disease outbreaks to determine the host genetic influence underlying infectious disease, current quantitative genetics methodology must be modified to take transmission dynamics into account. In quantitative genetic analyses, it is customary to assume that binary data is the realisation of a probability. Thus an important step is to identify the probability function that links the epidemiological parameters of interest, such as susceptibility and infectiousness, to the probability of becoming infected.

Therefore, the aim of this study was to derive an analytical expression for the probability of an individual to become infected within a given time period. We demonstrate how this can be achieved by integrating fundamental principles of epidemiology into the quantitative genetics framework. We then validate this analytical expression by comparing it with established theory in the case of homogeneous populations and by using simulated disease data generated for a range of epidemiological scenarios in genetically heterogeneous populations. Finally, we examine the implications for implementing this probability function into quantitative genetic analyses.

4.1 Methods

4.1.1 Epidemiological principles and approaches

The study of infectious diseases typically falls within the realm of epidemiology. A key measure in epidemiology is the basic reproduction number R_0, defined as the expected number of secondary infections that one infectious individual causes in an otherwise susceptible population (Anderson and May 2006). Efforts for

epidemiological control of infections are targeted to reduce R_0, ideally to a value below one, because if R_0 is less than one, infection is unlikely to spread and expected to die out. The higher R_0 is, the greater are the risk and severity of epidemics (Anderson and May 2006). This key definition points to two important host characteristics that control the spread of infection: first, the susceptibility of non-infected individuals, i.e. the propensity of becoming infected upon contact with an infectious individual or substance, and second, the infectiousness of the infected individuals, i.e. the ability of an infected individual to transmit the infection. As stipulated by Lloyd-Smith et al. (2006), for diseases transmitted by direct contact, infectiousness (or, using their terminology, individual reproductive number with population mean R_0) can be regarded as the product of three factors: *c*, the rate at which an infectious individual comes into contact with others in the population; *f*, the probability that the disease is transmitted to a susceptible individual, given contact; and *D*, the duration of the infectious period. All three components may harbour exploitable genetic variation.

Epidemiologists rely heavily on mathematical models of transmission dynamics to predict the outcome of control strategies. For instance, using a conventional compartmental *SIR* model that describes the transition of individuals between the Susceptible (*S*), Infected (*I*) and Recovered or Removed (*R*) compartment, the change in disease prevalence is described by $\frac{dI}{dt} = \beta S(t)I(t) - \gamma I(t)$ with parameters β (transmission coefficient) and γ (recovery rate) (Keeling and Rohani 2008). This differential equation represents infection as a dynamic process that arises from the interaction between susceptible and infected individuals (through the use of a multiplicative term in *S* and *I*). The transmission coefficient β is the product of the contact rate and the probability that the contact between an infectious and a susceptible individual results in a successful transmission (Keeling and Rohani 2008), and thus, depends on the susceptibility of the susceptible individual and the infectiousness of the infectious individual. Furthermore, for *SIR* models with constant population size, the limit of the probability *P(t)* of an initially susceptible individual to become infected within a time period *t* as $t \to \infty$ is given by

$$P(t) = 1 - e^{-\Lambda(t)} \quad (4.1)$$

Where Λ (t) = R_0 * R(t)/S_0 denotes the cumulative force of infection, i.e. the integral of the rate at which susceptible individuals become infected from time 0 to *t*, and *R(t)* and S_0 are the number of recovered individuals at time t and the initial number of susceptible individuals, respectively (Keeling and Rohani 2008).

Although epidemiologists acknowledge that there may be variation between individuals in both susceptibility and infectivity e.g. (Velthuis *et al.* 2003), classical epidemiology assumes homogeneity between individuals or within subgroups of individuals and therefore excludes the concept of host genetics. However, this gap has been shown to have a profound impact on the prediction of disease risk and prevalence, e.g. (Doeschl-Wilson *et al.* 2011; Nath *et al.* 2008; Springbett *et al.* 2003). In particular, recent field studies have elucidated the important role of super-spreaders, the small proportion of highly infectious individuals responsible for the majority of transmission events, on the occurrence and severity of disease outbreaks across a range of diseases (Chase-Topping *et al.* 2008; Lloyd-Smith *et al.* 2005; Matthews *et al.* 2009; Stein 2011). Note that super-spreaders confer host heterogeneity in infectiousness, not in resistance. Therefore, understanding and controlling heterogeneity in infectiousness, i.e. not only resistance, is now recognized as an important measure to control disease (Lloyd-Smith *et al.* 2005). However, to date, the genetic contribution of the host to this variation in infectiousness is unknown since genetic analyses tend to focus on disease resistance and, as demonstrated in chapters 2 & 3, fail to fully capture host genetic variation in infectiousness, if present, from binary disease data.

4.1.2 Derivation of a genetic-epidemiological probability function

Binary disease phenotypes can be considered as the realization of a probability of having the observed disease phenotype. In this section, we will extend the

epidemiological equation (4.1) for the (cumulative) probability of an individual to become infected by a time t for a heterogeneous host population with variation in both host susceptibility and infectiousness. For this purpose, we define f_k as the probability of an infectious individual k to infect a susceptible individual with unit susceptibility following contact, and g_j as the susceptibility of an individual j following contact with an infectious individual of unit infectivity. Furthermore, we define the indicator $X_{f,k}(t)$ to be equal to 1 if k is infectious at time t and to 0 otherwise. Then, the probability of a susceptible individual j of becoming infected following contact with individual k at time t is the product $g_j X_{f,k}(t) f_k$. Let c_{jk} be the expected number of contacts in a unit time interval between individuals j and k. Thus, following the same approach as in Keeling and Rohani (2008), for a susceptible individual not to become infected in a unit time interval, none of the contacts must result in infection. In other words, the probability of a susceptible individual j to avoid getting infected in a unit time interval is equal to

$$\prod_{k=1,k\neq j}^{n} \left(1 - g_j X_{f,k}(t) f_k\right)^{c_{jk}}. \tag{4.2}$$

The probability $P_j^*(\delta t)$ of a susceptible individual j to become infected during a sufficiently short time interval $[t, t + \delta t]$ during which the infection status of infectious individuals does not change is therefore,

$$P_j^*(\delta t) = 1 - \left(\prod_{k=1,k\neq j}^{n} \left(1 - g_j X_{f,k}(t) f_k\right)^{c_{jk}} \right)^{\delta t}. \tag{4.3}$$

Let $P_j(t)$ be the probability of individual j, which was susceptible at time zero, to have become infected by time t. Then for a small time-step δt,

$$P_j(t + \delta t) = P_j^*(\delta t)\left(1 - P_j(t)\right) + P_j(t). \tag{4.4}$$

Note, that this equation may encompass single and repeated infections (e.g. infected, recovered and re-infected) within the time interval from 0 to *t*. Rearranging the above equation, dividing by δt and taking the limit $\delta t \to 0$ leads to

$$\frac{dP_j(t)}{dt} = \lim_{\delta t \to 0} \frac{P_j^*(\delta t)}{\delta t}\Big(1 - P_j(t)\Big). \tag{4.5}$$

Note that the expression for $P_j^*(\delta t)$ above can be written as

$$P_j^*(\delta t) = 1 - \exp\left(\delta t \sum_{k=1, k\neq j}^{n} c_{jk} \ln\big(1 - g_j X_{f,k}(t) f_k\big)\right). \tag{4.6}$$

Using the power series expansion of the exponential function, and dividing by δt and taking the limit $\delta t \to 0$, leads to

$$\begin{aligned} \lim_{\delta t \to \mathbf{0}} \frac{P_j^*(\delta t)}{\delta t} &= -\sum_{k=\mathbf{1}, k\neq j}^{n} c_{jk} \ln\big(1 - g_j X_{f,k}(t) f_k\big) \\ &\approx g_j \sum_{k=\mathbf{1}, k\neq j}^{n} c_{jk} X_{f,k}(t) f_k, \end{aligned} \tag{4.7}$$

using the approximation $\ln(1 - x) \approx -x$ for small x. Substituting this last expression into the differential equation (4.5) yields

$$\frac{dP_j(t)}{dt} = g_j \sum_{k=1, k\neq j}^{n} c_{jk} X_{f,k}(t) f_k \,(1 - P_j(t)). \tag{4.8}$$

Now, define

$$\Lambda_j(t) := \int_0^t \left(g_j \sum_{k=1, k\neq j}^{n} c_{jk} X_{f,k}(u) f_k \right) du. \tag{4.9}$$

so that

$$\frac{dP_j(t)}{dt} = \frac{d\Lambda_j(t)}{dt}\,(1 - P_j(t)). \tag{4.10}$$

Multiplying both sides of (4.10) by $e^{\Lambda_j(t)}$ and collecting all terms to the left hand side leads to

$$\frac{d}{dt}\left(\, e^{\Lambda_j(t)} P_j(t) - \, e^{\Lambda_j(t)}\right) \, = \, 0, \tag{4.11}$$

or

$$e^{\Lambda_j(t)}\left(P_j(t) - 1\right) = constant. \tag{4.12}$$

Hence, the solution of the differential equation (4.10) is

$$P_j(t) = 1 + \left(P_j(0) - 1\right)e^{-\Lambda_j(t)}. \tag{4.13}$$

The probability $P_j(0)$ can be estimated as the prevalence at the beginning of an observation period. For simplicity, however, from now on we will assume that $P_j(0) = 0$ and hence,

$$P_j(t) = 1 - e^{-\Lambda_j(t)}. \tag{4.14}$$

Note that the quantity $\Lambda_j(t)$ defined above can be written as

$$\Lambda_j(t) = \, g_j \sum_{k=1, k\neq j}^{n} c_{jk}\, f_k D_k(t). \tag{4.15}$$

where $D_k(t)$ denotes the duration of time within the interval $[0,t]$ during which individual k is infectious. Thus, if k has not become infected by time t, $D_k(t) = 0$, otherwise

$$D_k(t) = \sum_{i=1}^{m} \left(min\left(t_{E_i}, t\right) - t_{S_i} \right),$$

where *m* denotes the number of times that individual *k* got infected during [0,t] and t_{S_i} and t_{E_i} denote the start and end of the corresponding infectious periods, respectively.

4.1.3 Function validation

Two forms of validation of the above derived probability function given by equation (4.14) with $\Lambda_j(t)$ defined in (4.15) were carried out. First, we assessed whether in the extreme case of zero heterogeneity in susceptibility and infectiousness, the derived function is consistent with existing epidemiological theory. Second, the function was validated with binary disease data (infected or not infected) generated by simulated stochastic epidemics in closed genetically heterogeneous populations of constant size, as described in detail in chapters 2 & 3. Two methods were chosen to illustrate this second validation: (i) a direct comparison of the probability of infection predicted by the derived analytical expressions (4.14) and (4.15) with the proportion of individuals that became infected in the simulations, and (ii) Receiver Operating Characteristic (ROC) curves. A ROC curve is a widely used graphical representation of the ability of a predictor to discriminate between cases and controls by plotting the True Positive Rate (TPR = sensitivity) against the False Positive Rate (FPR = 1-specificity) (Heagerty and Zheng 2005). Here, the ROC curves plot the proportion of infected individuals that have an estimated probability of infection greater than a given threshold (True Positives) against the proportion of non-infected individuals that have an estimated probability of infection greater than this same threshold (False Positives). Thus, the Area Under this Curve (AUC) describes the probability of correctly ranking any infected/non-infected pair of individuals using the derived probability function. Thus, if the analytical prediction is entirely unrelated to the probability of becoming infected in the simulations, then individuals would be classified at random and the AUC would be equal to 0.5. However, if our function accurately describes the probability of becoming infected in the simulations, then the

AUC would be close but not equal to 1, due to the stochastic nature of the simulations.

The stochastic epidemiological model used for validation simulates disease progression in isolated groups of n individuals and provides the disease status of individuals (infected / not infected) over time as output. The epidemic was simulated as a Poisson process, starting with one randomly chosen infected individual per group. The times at which subsequent infection and recovery events occurred and which individuals were affected were determined by the pairwise transmission parameters $\beta_{jk}(t)$ and by the recovery rates $\gamma_j(t)$, respectively, as outlined below. It was assumed that infected individuals became immediately infectious and remained infectious until they recovered. No transmission was assumed between groups.

Individual variation in host susceptibility and infectiousness was first incorporated into the model by assigning for each individual j its own level of susceptibility g_j and infectivity f_j. The dynamic, pairwise transmission parameter $\beta_{jk}(t)$ was then calculated as:

$$\beta_{jk}(t) = -c_{jk} ln\big(1 - X_{g,j}(t) g_j X_{f,k}(t) f_k\big), \quad (4.16)$$

as derived in Chapter 2. Thus, in line with standard epidemiological theory $\beta_{jk}(t)$ encapsulates the contact rate and the transmission probability. To reflect whether susceptibility and infectivity are expressed at time t, the individual constants g_j and f_k are scaled by $X_{g,j}(t)$ and $X_{f,k}(t)$, respectively, which are equal to 1 if j is susceptible at time t and if k is infectious at time t, respectively, and 0 otherwise. Similarly, individual recovery rates were assumed to be equal to $\gamma_j(t) = X_{f,j}(t)\gamma_j$, with γ_j and $X_{f,j}(t)$ as defined above.

It was initially assumed that host susceptibility and infectivity were the only sources of individual variation. Thus, parameter γ_j was set equal to 0.1 for all individuals. For simplicity, it was further assumed that the expected number of contacts per unit time interval between two individuals in the same group was homogeneous and, without

loss of generality, was set equal to $c_{jk} = 1$. This homogeneity assumption is likely to be satisfied in intensive farming conditions. The values of $\beta_{jk}(t)$ and $\gamma_j(t)$ were calculated at each event time, starting from time zero. Based on these, Gillespie's direct algorithm was used to determine the next event (infection or recovery), the time of the event and the affected individuals, as outlined in Chapter 2. The simulation was run until the time t at which approximately 50% of individuals had become infected.

In order to demonstrate that the derived probability function given by equations (4.14) and (4.15) is valid for a range of epidemiological models, binary disease data were also generated by simulating an epidemic using a stochastic SIR model with additional variation in recovery rate γ and a stochastic SLIRS model, following the same principles as described above. In a SLIRS model, the epidemiological compartments are: Susceptible (S), Latently infected but not infectious (L), Infectious (I), Recovered and temporarily immune (R), and Susceptible (S). The speed of transition between compartments S and L is given by $\beta_{jk}(t)$, as described above. Similarly, all other individual transition speeds were assumed equal to a constant value for individuals in the relevant compartment and 0 otherwise. Specifically, the constants were; 0.5 for L → I, 0.1 for I → R and 0.2 for R → S. Similar to the previous simulation, it was assumed that the expected number of contacts between two individuals per time unit $c_{jk} = 1$ for all individuals from the same group. This simulation was run until the same value of t as above, which resulted in approximately 58% of individuals becoming infected.

Thus, the different epidemiological models used for simulation were (i) a SIR model with host variation in susceptibility and infectivity only; (ii) a SIR model with host variation in susceptibility, infectivity and recovery rate; and (iii) a SLIRS model with host variation in susceptibility and infectivity only.

Each model was run for a population of size N = 100 000 individuals, randomly divided into 10 000 isolated groups of size 10 chosen, which is equivalent to simulating 10 000 independent epidemics. Susceptibility and infectivity were

assumed to be distributed according to a right-skewed gamma distribution $\Gamma(a,\theta)$, which is representative for a variety of infectious diseases (Lloyd-Smith *et al.* 2005). Moreover, skewed distributions allow for larger variation when the distribution is confined to positive values. For simplicity, susceptibility and infectivity were assumed to be independent. Similarly, additional individual variation in recovery rate was incorporated into the above described SIR model by sampling individual time to recovery $1/\gamma_j$ from a right-skewed Gamma distribution $\Gamma(2,5)$. In other words, it was assumed that most individuals recover quickly, that a few individuals may take a very long time to recover, and that the mean time to recovery was ten time units. This simulation was run until the same value of t as above, which resulted in approximately 41% of individuals to become infected.

Each epidemiological model provided the binary disease state (infected/ not infected by time t) for every individual as output. Furthermore, the period of time during which each individual remained infectious (D_k) was recorded for validation purposes. Note that the duration of the infectious period D in equation (4.15) captures individual variation in the transmission speeds between compartments L→I, I →R and R→S. Knowledge of the infectious period, together with the known input values of c, g and f, allowed calculation of the quantity $\Lambda_j(t)$ using equation (4.15) and hence the probability of becoming infected by a time t, based on equation (4.14). This was then compared with the observed proportion of individuals that became infected by time t in the simulations, within a given class of $\Lambda_j(t)$. The class size for $\Lambda_j(t)$ was taken as 0.02 to ensure that sufficient records were available within each class.

4.2 Results

4.2.1 Validation of the probability function

4.2.1.1 Concordance with epidemiological theory

We first demonstrate that for homogeneous populations, equations (4.14) and (4.15) are consistent with existing epidemiological theory and with the method of survival analysis. In a homogeneous population, i.e. when there is no variation in susceptibility ($g_j = g$ for each individual j), infectivity ($f_k = f$ for all k), contact rate

($c_{jk} = c$ for all j, k) or any of the other epidemiological parameters, equation (4.15) becomes

$$\Lambda_j(t) = \Lambda(t) = cgf \sum_{k=1, k\neq j}^{n} D_k(t). \quad (4.17)$$

Also, following equation (4.16), in the case of homogeneity, for any pair consisting of a susceptible individual j and an infectious individual k (i.e. $X_{g,j}(t) = X_{f,k}(t) = 1$), the transmission coefficient is

$$\beta = -c.ln(1 - gf) \approx cgf, \quad (4.18)$$

for small values of g and f.

Furthermore, the sum of the infectious period of each individual in a group, within the time interval from 0 to t, can be written as

$$\sum_{k=1, k\neq j}^{n} D_k(t) = \int_0^t I(\tau)d\tau, \quad (4.19)$$

where $I(\tau)$ denotes the number of infectious individuals at time τ. In an SIR model with constant recovery rate γ, the number of recovered individuals, R, changes over time according to $dR/dt = \gamma I(t)$, thus yielding the following for the above sum over infectious periods

$$\sum_{k=1, k\neq j}^{n} D_k(t) = \frac{1}{\gamma} R(t). \quad (4.20)$$

Note that in an SIR model, the basic reproductive ratio R_0 is

$$R_0 = \beta \frac{S_0}{\gamma}, \quad (4.21)$$

where S_0 is the number of susceptible individuals at the start of the epidemic (Keeling and Rohani 2008). Substituting equations (4.18) to (4.21) into (4.17), yields for $\Lambda_j(t) = \Lambda(t)$

$$\Lambda(t) \approx \frac{R_0\, R(t)}{S_0}, \tag{4.22}$$

and hence for $P_j(t) = P(t)$ according to equation (4.14)

$$P\,(t) \approx 1 - exp\left(-\frac{R_0\, R(t)}{S_0}\right).$$

Hence, the expression for the probability of becoming infected derived in section 4.1.2 for heterogeneous populations, i.e. equation (4.14), is consistent with equation (4.1) from epidemiological literature if there is no individual variation.

The probability function (4.14) is also consistent with the notion of failure in survival analysis, where the failure function F(t) represents the probability of failure by time *t* and is defined as $F(t) = 1 - e^{-\Lambda(t)}$, where $\Lambda(t)$ is the cumulative hazard function (Kalbfleisch and Prentice 2002). In this context, failure represents becoming infected. Therefore, equation (4.14) can be considered a failure function with a cumulative hazard function given by equation (4.15).

4.2.1.2 Function validation with simulated disease data

Figure 4-1 shows the proportion of individuals that had become infected by time *t* in the epidemiological simulations, for a given time *t* and calculated values of $\Lambda_j(t)$, as well as the analytical expression for the probability of becoming infected derived in equations (4.14) and (4.15). Figures 4-1a, b and c indicate that the probability function provides a good fit to the probability of becoming infected. Moreover, this function provides a robust fit across a range of epidemiological scenarios, as shown in Figures 4-1a, b and c for, respectively, the SIR model with variation in susceptibility and infectivity, with additional variation in recovery rate, and the

SLIRS model. Note that parameter values used in the simulations (see section 4.1.3.) are arbitrary and not expected to affect the fit.

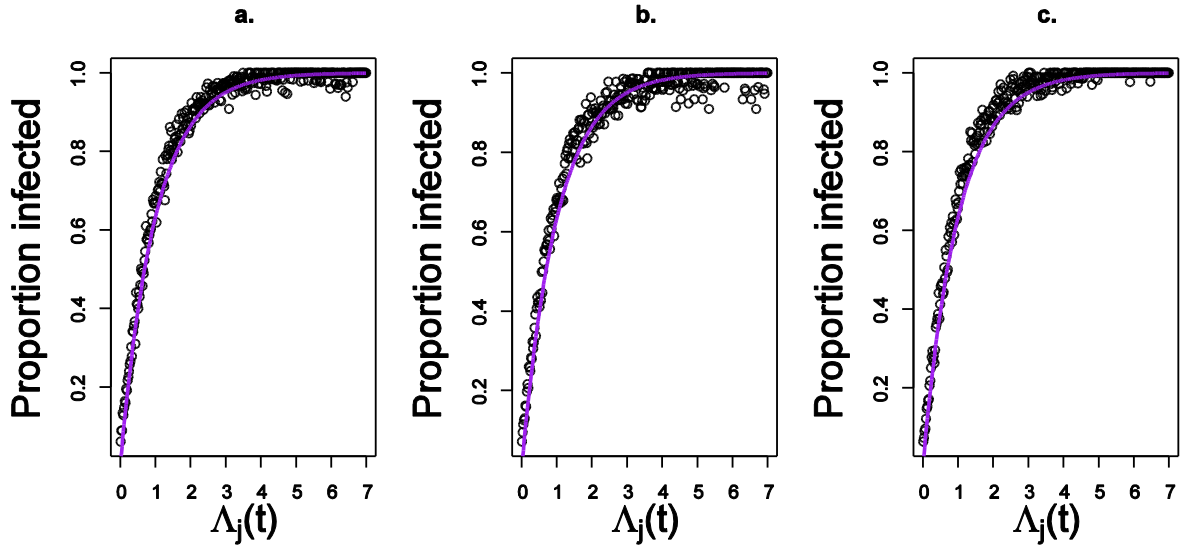

Figure 4-1 Comparison of the probability function (equations (4.14) and (4.15)) with results from simulated disease data

For details regarding simulation parameters see section 4.1.3.; data points: proportion of infected individuals for a given class of $\Lambda_j(t)$ using equation (4.15) with class size 0.02; curve: expected probability of becoming infected by time *t* following equations (4.14) and (4.15); panels: **a**. SIR model with variation in susceptibility and infectivity only, **b**. SIR model with variation in recovery rate, and **c**. SLIRS model.

Figure 4-2 shows ROC curves for predicting whether an individual has become infected or not by time *t*, with the derived probability given by equations (4.14) and (4.15) as the classification criterion. According to Figure 4-2, the derived probability is effective at predicting whether an individual will become infected or not by time *t*, in a manner that is consistent with an accurate probability function, i.e. with an AUC that is close to, but not equal to, 1. Moreover, the predictive ability of the derived probability function is robust across a range of epidemiological scenarios, with an AUC between 96-97% for all simulations.

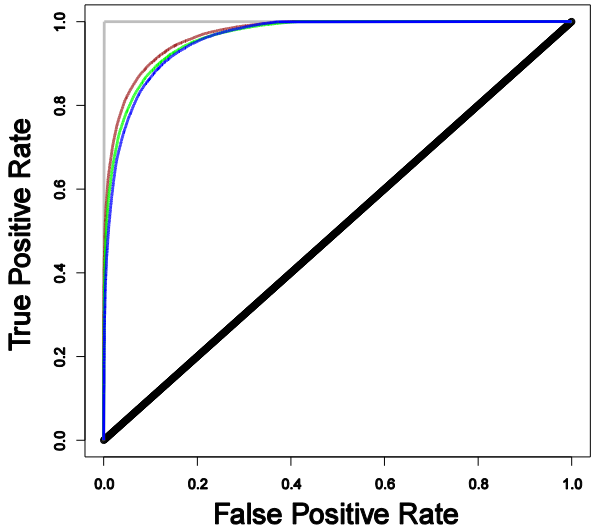

Figure 4-2 ROC curves for predicting disease status using the probability function (equations (4.14) and (4.15))

Curves: green = data from simulation of the SIR model with variation in susceptibility and infectivity (AUC = 0.964); **blue** = data from simulation of SIR model with variation in susceptibility, infectivity and recovery rate (AUC = 0.960); **brown** = data from simulation of SLIRS model with variation in susceptibility and infectivity (AUC = 0.970); **black** = random classification (AUC = 0.5); **grey** = perfect classification (AUC = 1).

The probability function (4.14), with $\Lambda(t)$ defined in (4.15), captures different sources of host (genetic) variation, which may not be easy to estimate in practice. In particular, whereas susceptibility g and infectivity f may harbour substantial genetic variation, the duration of the infectious period D within a given time interval is more likely to depend upon a combination of various genetic (e.g. g, f and also in γ) and environmental (e.g., choice of time interval), or other stochastic factors. In order to determine the importance of estimating these components of $\Lambda_j(t)$ for predicting the future disease status of an individual, ROC curves were also generated with the classification criterion estimated by assuming either no (genetic) heterogeneity in g and f (i.e. calculating $\Lambda_j(t)$ according to equation (4.17)), or by assuming genetic heterogeneity but equal non-dynamic exposure ($D_k(t) = \bar{D}$ for each individual k) in the probability function. The first scenario may be considered to be in line with current epidemiological theory, as outlined in section 4.1.1. and equation (4.17),

whereas the second scenario may be considered to be more in line with current quantitative genetics theory that ignores dynamic exposure. Note that exact values of $D_k(t)$ may not be available from field data and, therefore, using the further approximation from equation (4.20) is more in line with current epidemiological practice. However, applying this approximation results in discrete values of $D_k(t)$ rather than a continuous curve (results not shown). Nonetheless, the resulting discrete values are close to the curve obtained without using this approximation. Figure 4-3 shows a comparison of the ROC curves that correspond to these 'epidemiological' and 'genetic' assumptions, with the ROC curve that combines genetics and epidemiology in the derived expression for $\Lambda_j(t)$ outlined in equation (4.15). The ROC curves in Figure 4-3 reveal that quantifying the exposure over time explains most of the ability to predict whether an individual will become infected or not. Furthermore, predictions of an individual's disease status are considerably improved when all sources of genetic and epidemiological variation are included in the calculations.

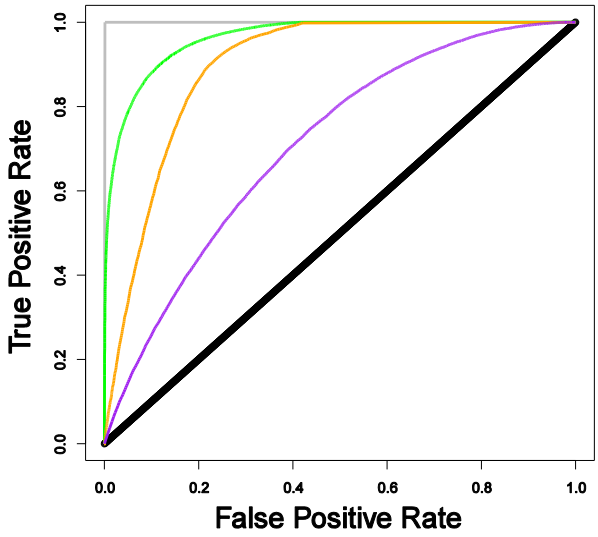

Figure 4-3 Effect of including different sources of host variation on the prediction of individual disease status

ROC curves calculated with data from simulation of the SIR model with variation in susceptibility and infectivity; the classification criterion used was the probability function equation (4.14) with $\Lambda_j(t)$ including different sources of variation; **Curves**: **green** = 'Genetic epidemiology' - $\Lambda_j(t)$ includes all sources of variation and was estimated based on equation (4.15) (AUC = 0.964); **orange** = 'Epidemiology' - $\Lambda_j(t)$ was estimated assuming no (genetic) variation in susceptibility and infectivity, as in equation (4.17) (AUC = 0.895); **purple** = 'Genetics' - $\Lambda_j(t)$ was estimated assuming (genetic) variation in susceptibility and infectivity, but equal non-dynamic exposure, i.e. $\boldsymbol{D_k(t) = \bar{D}}$ for each individual *k* (AUC = 0.710); **black** = random classification (AUC = 0.5); **grey** = perfect classification (AUC = 1).

4.3 Discussion

4.3.1 Extension to current epidemiological and quantitative genetics theories

Using mathematical principles, a genetic – epidemiological probability function was derived that links binary disease data to the underlying epidemiological traits, host susceptibility and infectiousness. The function is an extension of the established epidemiological equation for the probability of becoming infected by a time *t* (4.1) from homogeneous to heterogeneous populations. Indeed, in line with

epidemiological theory, the quantity $\Lambda_j(t)$ described in equation (4.15) may be called *the individual cumulative force of infection* of an individual *j* at time t. Defining infectiousness of individual *k* towards individual *j* until time *t* as the product $\phi_{jk}(t) = c_{jk}f_kD_k(t)$, as previously postulated by Lloyd-Smith et al. (2006), simplifies the expression for $\Lambda_j(t)$ to:

$$\Lambda_j(t) = g_j \sum_{k=1,k\neq j}^{n} \phi_{jk}(t). \quad (4.23)$$

Thus, the cumulative force of infection for an individual *j* is the product of the individual's susceptibility and the cumulative infectiousness of its group members towards it, which reflects that an infectious disease results from interactions between susceptible and infectious individuals. Note that under the assumption that $c_{jk} = c_k$ for each individual *k*, the infectiousness $\phi_{jk}(t)$ derived here corresponds to the individual reproductive number with population mean R_0, as defined in epidemiological literature (Lloyd-Smith *et al.* 2006). In the context of quantitative genetics, the cumulative infectiousness replaces the concept of exposure. Rather than an equal, constant and purely environmental exposure, as is typically assumed (Bishop *et al.* 2012), the individual cumulative force of infection in equation (4.23) illustrates that exposure depends on the number of infectious individuals, which may change over time as their infection status changes, as well as on their contact behaviour and infectivity, where some or all of these components may be partly genetically determined. In particular, the time $D_k(t)$ during which an individual remains infected may be partly genetically determined since it encapsulates several mechanisms that are determined by the immune system, such as recovery and latency. Thus, there is potentially much to be gained by incorporating epidemiological information into genetic analyses, and vice-versa, as illustrated in Figure 4-3.

The concept that an individual's phenotype is not only controlled by its own genes but also by the genes of interacting individuals is not new in quantitative genetics, and has already been successfully incorporated in the form of indirect (or associative) genetics effect (IGE) models (Bijma *et al.* 2007a; Bijma *et al.* 2007b;

Muir and Craig 1998). In chapters 2 & 3 we applied such IGE models to estimate genetic parameters associated with host susceptibility and infectivity from simulated binary disease data, and found that IGE models can indeed capture some of the genetic variation underlying infectiousness. However, in Chapter 3 we also found that use of the current IGE framework in the context of infectious disease has shortcomings since crucial dynamic aspects are ignored, which leads to bias in parameter estimates. As outlined in more detail below, the derived genetic-epidemiological probability function offers a means to extend the current IGE model framework to infectious diseases in populations that display genetic variation in diverse epidemiological traits for which expression varies throughout the time course of infection.

4.3.2 Implementation of the probability function into quantitative genetic analysis

In order to incorporate susceptibility and infectiousness into genetic selection programs, knowledge of the respective genetic (co)variances is required. Moreover, it might be desirable to use estimated breeding values of these traits for genetic selection or for genome-wide association studies. Estimation of breeding values by best linear unbiased prediction requires not only knowledge of the genetic variance (Falconer and Mackay 1996) but also the use of mixed models, as these allow simultaneous estimation of fixed effects and random genetic effects (Falconer and Mackay 1996). Susceptibility and infectiousness are difficult to measure directly and, as was assumed in this chapter, field disease data is often binary, indicating whether an individual became infected or not. It is customary to use a generalized linear (mixed) model (GL(M)M) to analyse binary or categorical data (Mccullagh and Nelder 1995). In such models, the observed trait is linked to an assumed linear model of the underlying continuous trait(s) via a non-linear link function. Canonical link functions that are commonly used for binary data are the probit and logit link functions (Mccullagh and Nelder 1995), which assume that the probability of the trait to be equal to one, i.e. to have become infected in our case, follows a cumulative normal or a logistic distribution, respectively (Mccullagh and Nelder 1995). Despite their convenient mathematical properties, neither distribution, however, arises

naturally from epidemiological theory, as demonstrated in the present study. A consequence of this is that interpretation of such analyses in terms of epidemiological parameters is problematic at best. A suitable link function for a GL(M)M transforms the observed trait into a linear expression of the parameters of interest. However, in the genetic epidemiological probability function $P_j(t)$ (equation (4.14) with $\Lambda_j(t)$ defined in equation (4.23)), the parameters of interest, i.e. the epidemiological traits susceptibility and infectiousness, enter in a multiplicative rather than in a linear manner. However, if there was genetic variation in susceptibility *only*, it follows from equations (4.14) and (4.23) that the probability $P_j(t)$ can be linked to the following linear model in susceptibility using a complementary log-log link function:

$$\ln\left(\Lambda_j(t)\right) = \ln(g_j) + \ln\left(\sum_{k=1,k\neq j}^{n} \phi_{jk}(t)\right). \qquad (4.24)$$

Assuming no genetic variation in the epidemiological traits c_{jk}, f_k and D_k that underly infectiousness, the second summand of equation (4.24) can be considered to be an error term $e_j(t)$. However, in contrast to using the canonical logit and probit link functions, this model captures and completely separates the individual's susceptibility from the dynamic aspects of exposure.

However, when there is genetic variation in both susceptibility and infectiousness, it is not straightforward to link the probability $P_j(t)$ of becoming infected to a linear model that includes both susceptibility and infectiousness. Indeed, the complementary log-log link function (4.24) is no longer adequate when there is variation in infectiousness since the logarithm of a sum does not equal the sum of the logarithms. It is, however, possible to linearize the cumulative force of infection from equation (4.23), in both susceptibility and infectiousness, using e.g. the Taylor series expansion of $\Lambda_j(t) = g_j \sum_{k=1,k\neq j}^{n} \phi_k(t)$ near the population mean susceptibility $\bar{g}$ and the population mean infectiousness $\bar{\phi}(t)$ up to time *t*:

$$\begin{aligned}\Lambda_j(t) = {} & (n-1)\bar{g}\,\bar{\phi}(t) + (n-1)\bar{\phi}(t)\left(g_j - \bar{g}\right) \\ & + \bar{g}\sum_{k=1,k\neq j}^{n}\left(\phi_{jk}(t) - \bar{\phi}(t)\right) \\ & + \left(g_j - \bar{g}\right)\sum_{k=1,k\neq j}^{n}\left(\phi_{jk}(t) - \bar{\phi}(t)\right)\end{aligned} \tag{4.25}$$

Note that the Taylor series of $\Lambda_j(t)$ in equation (4.25) is not truncated and that it includes only one non-linear term in susceptibility and infectiousness. Following a GL(M)M framework, if the last term of equation (4.25) was negligible, the expression for $\Lambda_j(t)$ would be linear and thus an appropriate link between observed binary disease data (infected or not infected) and the underlying epidemiological traits, host susceptibility and infectiousness.

Note that truncating equation (4.25) after the linear terms in g_j and $\phi_{jk}(t)$ corresponds to an IGE model for the individual cumulative force of infection $\Lambda_j(t)$. IGE models describe the phenotype P_j (here $P_j = \Lambda_j(t)$) of an individual *j* as a linear combination of the individual's direct effect P_{Dj}, and the cumulate indirect (or associative) effect P_{Sk} of its group members, i.e.

$$P_j(t) = \mu + P_{D_j} + \sum_{k=1,k\neq j}^{n} P_{S_k}, \tag{4.26}$$

with an underlying genetic component for both the direct and indirect effects and with μ denoting the population mean phenotype, e.g. (Bijma *et al.* 2007a; Bijma *et al.* 2007b). The connection between host infectiousness and indirect effects has been established previously in Chapter 2 but the exact nature of this connection was unknown. Thus, comparison of the linear part of equation (4.25) with equation (4.26) offers a new interpretation of direct and indirect effects in this context and of previous results. Indeed, according to equation (4.25), the direct effect corresponds to the susceptibility of individual *j* (expressed as deviation from the population mean susceptibility), scaled by the cumulative average infectiousness of the group

members up to time *t*, and the indirect (or associative) effect of a group member corresponds to its infectiousness (expressed as deviation from the population mean infectiousness until time *t*), scaled by the average population susceptibility. Furthermore, equation (4.25) may shed some light on potential causes for the bias observed in Chapter 3 in the genetic parameter estimates in infectivity. This bias may have resulted from the inadequacy of the linear and logit models used in the previous analyses, as neither emerges from epidemiological theory and the appropriate link function was yet unknown. Furthermore, as illustrated in equation (4.25), the non-linear interaction between susceptibility and infectiousness may become non-negligible if there are large deviations in infectiousness ϕ from the population mean. This is illustrated in Figure 4-4, which shows the ROC curves with the classification criterion estimated with the full (AUC = 0.964) and truncated (AUC = 0.751) versions of equation (4.25). In other words, in the presence of super-spreaders, i.e. highly infectious individuals, the use of a GL(M)M or any other linear framework is likely to create bias. For the purpose of identifying super-spreaders, it would therefore be desirable to develop computational algorithms that do not require linear approximations of the cumulative force of infection function. Such non-linear algorithms would also be needed to disentangle the individual components of infectiousness, e.g. to separate genetic variation in the ability to transmit the infection upon exposure (i.e. variation in *f*) from genetic variation in the duration of the infectious period (i.e. variation in *D*). These sources of variation likely correspond to different immunological processes (e.g. shedding vs. recovery) and may therefore be controlled by different sets of genes. However, separating infectiousness components in genetic analyses may come with additional data requirements. For example, repeated binary measurement of an individual's disease status over time rather than one single snapshot in time may be required to infer genetic variation in the duration of the infectious period. These measurements may be taken from on-going epidemics by using equation (4.13) instead of (4.14), with $P_j(0)$ equal to the prevalence of the disease in the first observation. Markov Chain Monte Carlo methods (Hastings 1970), with their hierarchical iterative sampling process, appear well suited to incorporate the dynamic expression of host susceptibility and infectiousness. Such methods may also lend themselves more easily to the consideration of other

uncertainties that frequently affect observed disease phenotypes, such as incomplete sensitivity or specificity of diagnostic tests.

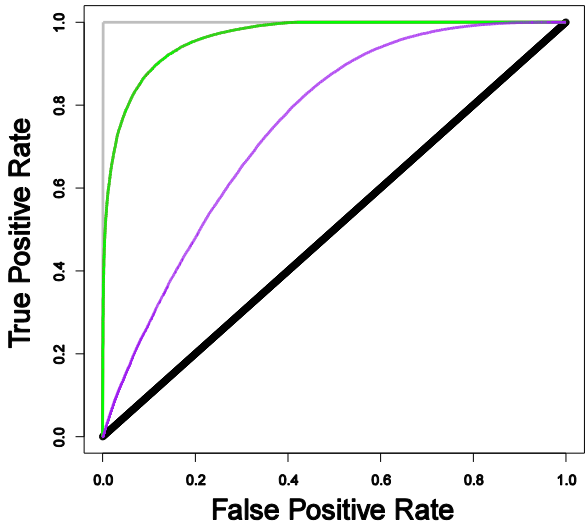

Figure 4-4 ROC curve for predicting disease status using an IGE model

Data from simulation of the SIR model with variation in susceptibility and infectivity; **Curves: green** = the probability function with lambda estimated as in equation (4.15) used as classification criterion (AUC = 0.964); **brown** (overlapping with green curve) = the probability function with lambda estimated using the Taylor expansion from equation (4.25) used as classification criterion (AUC = 0.964); **purple** = an IGE model (equation (4.26)) used as classification criterion (AUC = 0.751); **black** = random classification; **grey** = perfect classification.

4.4 Conclusions

We have derived a genetic epidemiological function for quantitative genetic analyses of binary infectious disease data that takes genetic variation and the dynamic expression of host infectiousness into account. The function describes the probability of an individual to become infected given its own susceptibility and the infectiousness of its group mates. When variation is limited to host susceptibility, it is possible to estimate genetic variation for this trait in a manner compatible with epidemiological dynamics using the complementary log-log link function. When

there is genetic variation in both susceptibility and infectiousness, it is possible to use the logarithmic link function with a linear IGE model but this is likely to generate prediction bias if there is a large variation in infectiousness. Future work will concentrate on developing computational algorithms that can incorporate the genetic epidemiological function without linear approximations, in order to identify potential genetic super-spreaders. These algorithms would enable us to uncover the genetics underlying epidemics and thus shape the epidemics of tomorrow.

4.5 References

Anderson, R. M., and R. M. May, 2006 *Infectious Diseases of Humans*. Oxford University Press, Oxford.

Biffani, S., S. Botti, A. Caprera, E. Giuffra and A. Stella, 2010 Genetic Susceptibility to Porcine Reproductive and Respiratory Syndrome (PRRS) virus in commercial pigs in Italy., pp. in *9th WCGALP*, Leipzig.

Bijma, P., W. A. Muir and J. a. M. Van Arendonk, 2007a Multilevel selection 1: Quantitative genetics of inheritance and response to selection. Genetics **175:** 277-288.

Bijma, P., W. M. Muir, E. D. Ellen, J. B. Wolf and J. a. M. Van Arendonk, 2007b Multilevel selection 2: Estimating the genetic parameters determining inheritance and response to selection. Genetics **175:** 289-299.

Bishop, S. C., A. B. Doeschl-Wilson and J. A. Woolliams, 2012 Uses and implications of field disease data for livestock genomic and genetics studies. Front Genet **3:** 114.

Bishop, S. C., and J. A. Woolliams, 2010 On the Genetic Interpretation of Disease Data. PLoS ONE **5:** e8940.

Boddicker, N., E. H. Waide, R. R. R. Rowland, J. K. Lunney, D. J. Garrick *et al.*, 2012 Evidence for a major QTL associated with host response to Porcine Reproductive and Respiratory Syndrome Virus challenge. Journal of Animal Science **90:** 1733-1746.

Chase-Topping, M., D. Gally, C. Low, L. Matthews and M. Woolhouse, 2008 Super-shedding and the link between human infection and livestock carriage of Escherichia coli O157. Nature Reviews Microbiology **6:** 904-912.

Doeschl-Wilson, A. B., R. Davidson, J. Conington, T. Roughsedge, M. R. Hutchings *et al.*, 2011 Implications of Host Genetic Variation on the Risk and

Prevalence of Infectious Diseases Transmitted Through the Environment. Genetics **188:** 683-693.

Falconer, D. S., and T. F. C. Mackay (Editors), 1996 *Introduction to Quantitative Genetics*. Pearson Education Limited, Harlow.

Hastings, W. K., 1970 MONTE-CARLO SAMPLING METHODS USING MARKOV CHAINS AND THEIR APPLICATIONS. Biometrika **57:** 97-&.

Heagerty, P. J., and Y. Y. Zheng, 2005 Survival model predictive accuracy and ROC curves. Biometrics **61:** 92-105.

Kalbfleisch, J. D., and R. L. Prentice, 2002 *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Keeling, M. J., and P. Rohani, 2008 *Modelling Infectious Diseases in Humans and Animals*. Princeton University Press, Princeton

Lloyd-Smith, J. O., S. J. Schreiber and W. M. Getz, 2006 Moving beyond averages: Individual-level variation in disease transmission, pp. 235-258 in *Mathematical Studies on Human Disease Dynamics: Emerging Paradigms and Challenges*, edited by A. B. GUMEL. American Mathematical Society, Providence.

Lloyd-Smith, J. O., S. J. Schreiber, P. E. Kopp and W. M. Getz, 2005 Superspreading and the effect of individual variation on disease emergence. Nature **438:** 355-359.

Matthews, L., R. Reeve, M. E. J. Woolhouse, M. Chase-Topping, D. J. Mellor *et al.*, 2009 Exploiting strain diversity to expose transmission heterogeneities and predict the impact of targeting supershedding. Epidemics **1:** 221-229.

Mccullagh, P., and J. A. Nelder, 1995 *Generalized Linear Models*. University Press, Cambridge.

Muir, W. M., and J. V. Craig, 1998 Improving animal well-being through genetic selection. Poultry Science **77:** 1781-1788.

Nath, M., J. A. Woolliams and S. C. Bishop, 2008 Assessment of the dynamics of microparasite infections in genetically homogeneous and heterogeneous populations using a stochastic epidemic model. Journal of Animal Science **86:** 1747-1757.

Springbett, A. J., K. Mackenzie, J. A. Woolliams and S. C. Bishop, 2003 The contribution of genetic diversity to the spread of infectious diseases in livestock populations. Genetics **165:** 1465-1474.

Stein, R. A., 2011 Super-spreaders in infectious diseases. International Journal of Infectious Diseases **15:** e510-e513.

Velthuis, A. G. J., M. C. M. De Jong, E. M. Kamp, N. Stockhofe and J. H. M. Verheijden, 2003 Design and analysis of an Actinobacillus pleuropneumoniae transmission experiment. Preventive Veterinary Medicine **60:** 53-68.

Chapter 5. An MCMC algorithm to estimate breeding values in susceptibility and infectivity from sequential binary disease data

We have shown in chapters 2 & 3 that it is possible to capture some genetic variation in infectivity if present, with an indirect genetic effects model as described by Muir for production traits (Muir 2005). However, this method was limited because it applied a static linear model to an essentially dynamic non-linear process as was shown in Chapter 4. Moreover, for reasons of feasibility this model was based on cross-sectional binary disease data which do not lend themselves to capture the dynamic nature of disease transmission.

In Chapter 4 we derived an analytical expression for the probability of an individual to become infected by a given time, which takes the non-linear interaction between susceptibility and infectivity and the disease dynamics into account. Here we develop a Markov Chain Monte Carlo (MCMC) algorithm that implements this expression in order to estimate breeding values in susceptibility and infectivity. We evaluate this algorithm by comparing the true simulated breeding values with the estimated breeding values, obtained by applying the algorithm to longitudinal binary disease data generated from epidemiological simulations with known variances in host infectivity and susceptibility.

5.1 Methods

5.1.1 Data requirements and assumptions

Genetic analyses usually require large sample sizes rendering data collection costly and time consuming and therefore every effort is taken to simplify and streamline this process. The simplest trait which may be gathered is a single binary snapshot, e.g. based on readily available PCR or ELISA assays which indicates whether an

individual became infected or not during a particular time. However, cross-sectional binary measurements, i.e. collected only once, would provide poor information about what is essentially a dynamic process as was seen in Chapters 2 & 3. For reasons of feasibility we therefore assume that the observed data would be repeated binary scores indicating whether individuals have become infected or not by specific sampling times. Moreover, Ødegård et al. have repeatedly demonstrated that it is possible to obtain more accurate estimates for survival breeding values using repeated binary scores than with cross-sectional binary data (Gitterle *et al.* 2006; Ødegård *et al.* 2006; Ødegård *et al.* 2007). Assuming a disease following an SI model, i.e. individuals are either susceptible (S) or infected (I), and accurate diagnostic tools, using repeated binary scores we would know that the time of infection τ_j of an individual *j* occurred during the transition period $[t_{B_j}, t_{E_j}]$ between the last sampling time t_{B_j} where individual *j* was known to be not infected and the first sampling time t_{E_j} where *j* was known to be infected (∞ if *j* didn't become infected). For example, if an individual is observed as not being infected on day two and infected on day five, then we know that the infection took place at some point between the second and fifth day. For diseases where other disease states occur, such as a latency or recovery period, the relationship between the observed data and the real time of infection is more complex. Nonetheless, the corresponding algorithms would be an extension of the one based on a simple SI model and hence as proof of concept we develop an algorithm based on the epidemiological SI model.

The algorithm is designed for natural field conditions with the infection spreading naturally through a population divided in a number of independent contact groups. However, it may also be used for challenge studies provided certain conditions are met. For example, given that we are interested in individuals' propensity to transmit the infection, challenging all individuals would be inappropriate. It is therefore important that each epidemic is started by a limited number of infected individuals in a group of otherwise naïve individuals. Moreover, housing related individuals in separate groups such that each group represents an independent epidemic is essential

for disentangling genetic from environmental effects. Moreover, knowledge of the pedigree or genomic information and group of each individual would be required.

For the methodology developed here, a difference between field conditions and a challenge study in the start of the epidemic must be taken into account. Indeed, in field conditions certain individuals will have been infected naturally prior to the first sampling time due to their susceptibility and the level of exposure. These individuals would simply be observed as having $t_B = -\infty$ and $t_E = 0$. In a challenge study, individuals are being artificially infected regardless of their genetic make-up and the binary data therefore provides no information on their susceptibility as they may not have become infected given natural conditions. These individuals therefore need to be treated differently. We refer to these artificially challenged individuals as the index cases and indicate their challenge status by an indicator variable s=0 and their binary observations are set as missing values with $t_B = -\infty$ and $t_E = +\infty$. All other individuals have challenge status s=1. The 'challenge' condition is therefore a special case of the 'field' condition.

5.1.2 Parameters

As in the previous chapter we define the susceptibility g_j of an individual *j* as the probability of a susceptible individual to become infected upon contact with an infectious individual with infectivity equal to one. We define infectivity f_j as the probability of an infected individual *j* to transmit an infection upon contact with a susceptible individual with susceptibility equal to one. As both traits are probabilities and therefore bounded by 0 and 1, for computational ease we assumed the existence of underlying traits which are the logit transformation $\psi = \ln(g) - \ln(1-g)$ of susceptibility and $\iota = \ln(f) - \ln(1-f)$ of infectivity, respectively. For simplicity, the population mean was assumed to be the only fixed effect. Hence, the distribution of ψ_j and ι_j , i.e. the uncertainty for these individual values, conditional on their respective breeding values and population means is equivalent to that of the environmental residual and is assumed to be distributed according to the following multivariate normal distribution,

$$\left[\begin{array}{c|cc} \psi_j & \mu_\psi, & a_{\psi_j} \\ \iota_j & \mu_\iota, & a_{\iota_j} \end{array}\right] \sim N\left(\left[\begin{array}{c} a_{\psi_j}+\mu_\psi \\ a_{\iota_j}+\mu_\iota \end{array}\right], \mathbf{V}\right). \quad (5.1)$$

Where a_{ψ_j} and a_{ι_j} are additive genetic effects, μ_ψ and μ_ι are population means, $\mathbf{V}$ is the environmental (co)variance matrix at the underlying level. Note that no assumption is made with regards to the distribution of susceptibility and infectivity at the population level.

The additive genetic effects are assumed to be distributed according to the following multivariate normal distribution,

$$\mathbf{a} \sim N\left(\mathbf{0}, \mathbf{A} \otimes \mathbf{G}\right). \quad (5.2)$$

Where **a** is a vector of all additive genetic effects**,** **A** is the FxF numerator relationship matrix (F = pedigree size) and **G** is the genetic (co)variance matrix of a_ψ and a_ι.

All known observations and unknown variables described above have been summarized in Tables 5-1 and 5-2 respectively.

Table 5-1 Summary of known values

$\mathbf{t}_B$	Vector (dimension *N*) of last individual sampling time with a susceptible status
$\mathbf{t}_E$	Vector (dimension *N*) of first individual sampling time with an infected status
s	Vector (dimension *N*) with 0 if artificially infected at t=0, 1 otherwise, for each individual
n	Group size(s)
N	Population size (with observed records)
F	*N* + number of known relatives without observed records
A	FxF numerator relationship matrix

Table 5-2 Summary of unknown variables

$\boldsymbol{\tau}$	Vector (dimension *N*) of times of infection of individuals in natural conditions
g_j	Susceptibility of individual *j*
$\boldsymbol{\psi}$	Vector (dimension *N*) of underlying susceptibility $\psi_j = \ln(g_j) - \ln(1 - g_j)$ for each individual *j*
f_j	Infectivity of individual *j*
$\boldsymbol{\iota}$	Vector (dimension *N*) of underlying infectivity $\iota_j = \ln(f_j) - \ln(1 - f_j)$ for each individual *j*
$\boldsymbol{a}_\psi$	Vector (dimension *F*) additive genetic effects for the underlying susceptibility
$\boldsymbol{a}_\iota$	Vector (dimension *F*) of additive genetic effects for the underlying infectivity
μ_ψ	Population mean underlying susceptibility
μ_ι	Population mean underlying infectivity
G	2x2 genetic (co)variance matrix
V	2x2 environmental (co)variance matrix

5.1.3 Probability density functions

As pointed out in the introduction to this chapter, the following methodology is based on the assumption of a SI model i.e. assuming that individuals become immediately infectious upon infection and do not recover. Therefore, the cumulative density function of an individual to become infected by a time *t*, derived in Chapter 4, takes the following form,

$$P(\tau_j \leq t) = 1 - \exp\left(-g_j \sum_{k=1,k\neq j}^{n} (f_k(t - \tau_k s_k)H(t - \tau_k s_k))\right). \quad (5.3)$$

Where H(x) represents the Heaviside step function, i.e. H(x)=1 when $x \geq 0$ and 0 otherwise and s_k = 0 if individual *k* was infected artificially prior to the start of the observation period and 1 otherwise. The sum is taken over all group mates of

individual *j*. The probability density function (p.d.f.) of an individual to become 'naturally' infected, at time *t*, is therefore equal to,

$$\begin{aligned} p(\tau_j = t) &= \frac{d}{dt} P(\tau_j \leq t) \\ &= g_j \sum_{k=1,k\neq j}^{n} (f_k H(t - \tau_k s_k)) \\ &\exp\left(-g_j \sum_{k=1,k\neq j}^{n} (f_k (t - \tau_k s_k) H(t - \tau_k s_k)) \right) \\ &= \frac{1}{1 + e^{-\psi_j}} \left(\sum_{k=1,k\neq j}^{n} \frac{H(t - \tau_k s_k)}{1 + e^{-\iota_k}} \right) \\ &\exp\left(-\frac{1}{1 + e^{-\psi_j}} \sum_{k=1,k\neq j}^{n} \frac{(t - \tau_k s_k) H(t - \tau_k s_k)}{1 + e^{-\iota_k}} \right) \end{aligned} \tag{5.4}$$

Following Bayes' Theorem, assuming flat priors for the genetic and environmental (co)variances and the population means, the p.d.f. of the unknown variables (Table 5-2) given the observed transition times t_B and t_E, is given by,

$$\begin{aligned} & p(\boldsymbol{\tau}, \mathbf{s}, \boldsymbol{\psi}, \boldsymbol{\iota}, \mathbf{a}_\psi, \mathbf{a}_\iota, \mu_\psi, \mu_\iota, \mathbf{G}, \mathbf{V} | \mathbf{t}_B, \mathbf{t}_E) \\ \propto & p(\mathbf{t}_B, \mathbf{t}_E | \boldsymbol{\tau}, \mathbf{s}, \boldsymbol{\psi}, \boldsymbol{\iota}, \mathbf{a}_\psi, \mathbf{a}_\iota, \mu_\psi, \mu_\iota, \mathbf{G}, \mathbf{V}) \\ & p(\boldsymbol{\tau}, \mathbf{s}, \boldsymbol{\psi}, \boldsymbol{\iota}, \mathbf{a}_\psi, \mathbf{a}_\iota, \mu_\psi, \mu_\iota, \mathbf{G}, \mathbf{V}) \\ = & p(\mathbf{t}_B, \mathbf{t}_E | \boldsymbol{\tau}) p(\boldsymbol{\tau} | \mathbf{s}, \boldsymbol{\psi}, \boldsymbol{\iota}) p(s) p(\boldsymbol{\psi}, \boldsymbol{\iota} | \mathbf{a}_\psi, \mathbf{a}_\iota, \mu_\psi, \mu_\iota, \mathbf{V}) \\ & p(\mathbf{a}_\psi, \mathbf{a}_\iota | \mathbf{G}) p(\mu_\psi) p(\mu_\iota) p(\mathbf{V}) p(\mathbf{G}) \\ \propto & p(\mathbf{t}_B, \mathbf{t}_E | \boldsymbol{\tau}) p(\boldsymbol{\tau} | \mathbf{s}, \boldsymbol{\psi}, \boldsymbol{\iota}) \\ & p(\boldsymbol{\psi}, \boldsymbol{\iota} | \mathbf{a}_\psi, \mathbf{a}_\iota, \mu_\psi, \mu_\iota, \mathbf{V}) p(\mathbf{a}_\psi, \mathbf{a}_\iota | \mathbf{G}). \end{aligned} \tag{5.5}$$

Where,

$$p(\mathbf{t}_B, \mathbf{t}_E | \boldsymbol{\tau}) p(\boldsymbol{\tau} | \mathbf{s}, \boldsymbol{\psi}, \boldsymbol{\iota}) =$$

$$\prod_{j=1}^{N} H\left(\tau_j - t_{B_j}\right) H\left(t_{E_j} - \tau_j\right) p(\tau_j | \psi_j, s_{-1}, \iota_{-1}). \tag{5.6}$$

Note that, for each individual only the time period where their infection status changes, from susceptible to infected, is considered. Moreover, it is assumed that individuals infected during the same time period do not affect each other. The density for τ_j is given by equation (5.4).

Given the multivariate normal distributions assumed in (5.1) & (5.2), the joint density of the underlying susceptibility and infectivity and their corresponding genetic and environmental parameters is given as,

$$\begin{aligned} & p(\boldsymbol{\psi}, \boldsymbol{\iota}|\mathbf{a}, \mu_\psi, \mu_\iota, \mathbf{V}) p(\mathbf{a}|\mathbf{G}) \\ =& \frac{|\mathbf{V}|^{-N/2}}{(2\pi)^N} \exp\left(-\frac{1}{2}\mathbf{y}^T(\mathbf{V}^{-1} \otimes \mathbf{I}_N^{-1})\mathbf{y}\right) \\ & \frac{|\mathbf{A}|^{-1}|\mathbf{G}|^{-F/2}}{(2\pi)^F} \exp\left(-\frac{1}{2}\mathbf{a}^T(\mathbf{G}^{-1} \otimes \mathbf{A}^{-1})\mathbf{a}\right). \end{aligned} \tag{5.7}$$

Where, **G** and **V** are the genetic and environmental (co)variance matrices, respectively, and

$$\mathbf{y} = \begin{pmatrix} \boldsymbol{\psi} - \boldsymbol{\mu}_\psi - \boldsymbol{a}_\psi \\ \boldsymbol{\iota} - \boldsymbol{\mu}_\iota - \boldsymbol{a}_\iota \end{pmatrix}.$$

Hence the full joint posterior distribution outlined in expression (5.5) is proportional to the product of equations (5.6) and (5.7).

5.1.4 Evaluation

In order to evaluate the above hierarchical framework, sequential binary disease data was generated. For this purpose epidemics were simulated with known genetic variation in host susceptibility and infectivity as inputs using a simulation, similar to that developed in Chapter 2, outlined below. The simulation outputs were then analysed with an MCMC algorithm using the probability density functions described above. Estimates of accuracy of this methodology were obtained by comparing the true values input in the simulation with the estimated values obtained by inference.

These estimates were used to evaluate first the program and its theoretical framework by sampling each parameter conditional on the true values of all other parameters. Then the required burn-in period and the acceptance rates associated with the proposal distributions were examined for the full stochastic sampling algorithm. The algorithm was then used with a range of simulations to examine factors including the effect of group size and infection time on the accuracy of parameter estimates. A more detailed account of these evaluation processes is given below.

5.1.4.1 Evaluation of the program and its theoretical framework

The Metropolis-Hastings (MH) MCMC algorithm was chosen for the implementation. The MH algorithm creates marginal densities for each variate by sampling from a proposal distribution and accepting or rejecting the samples relative to their probability density function conditional on all the other variates. This required deriving the conditional posterior distribution of each scalar parameter pertaining to the model from the joint posterior distribution outlined in equation (5.5). The conditional p.d.f. obtained when fixing all other variates to their true value from the simulation reflects the best case scenario where the marginal of all other variates are sharply peaked around their true values. We therefore sampled each variate from their conditional p.d.f. with all other variates, except for τ, equal to their true value in order to test whether it would be feasible in theory to infer the value of these variates from their marginals. The values of τ were always sampled as there is no known value for the index cases and the individuals which never became infected. Moreover, when individuals did become infected naturally τ_j is constrained between the relatively small interval $[t_{B_j}, t_{E_j}]$. Note that, unless stated otherwise, knowledge of parameter values is not assumed in all other evaluations.

5.1.4.2 Proposal distributions and burn-in period

In order to implement the algorithm all parameters were sampled for each individual from the full joint density outlined in equation (5.5) using the MH algorithm. The proposal distributions used for the individual variates were: flat between t_{B_j} and t_{E_j}

for the infection time τ_j, N~(-1,3.24) for the underlying susceptibility and infectivity and N~(0,1.21) for the corresponding breeding values. The proposal distribution for the underlying phenotypes were chosen such that 99% of the values sampled would be situated between [0.002, 0.988] and 50% between [0, 0.27] on the probability level. The proposal distributions for the population variates were: lnN~(previous estimate, 0.04) for the environmental and genetic variances, N~(previous estimate, 0.04) for the environmental and genetic covariances and N~(previous estimate, 0.01) for the underlying population means. As the previous estimate is used for the mean of the proposal distribution for the population parameters, the MH algorithm was repeated 100 times between each value of the chain to ensure that the values of the chain remain independent samples.

In order to identify the burn-in period, the algorithm was run with two different sets of starting values. The values chosen were two standard deviations either side of the proposed sampling distributions mean for the individual parameters or from the true values for the population parameters. The true values for the population parameters are given in section 5.1.4.3. below. The starting values used to identify the burn-in period as well as those used for subsequent analyses are given in Table 5-3.

Having taken care to discard the samples prior to a burn-in period (see section 5.2.2), the samples obtained with the MH algorithm were used to plot marginal density functions for each variate. Given that there are several variables per individual it would not be feasible to graphically represent every single one of them. In order to evaluate the algorithm at the preliminary phase, marginal densities were plotted for a subset of individuals using both histograms and boxplots. These individuals were chosen according to their infection time in order to assess the impact of infection time on the ability to infer each parameter. In particular, ten sires, ten index cases, ten individuals with the lowest infection time, ten individuals infected around the median infection time, ten individuals with the highest infection time and ten individuals which did not become infected were chosen. Care was taken that the individuals chosen for each category came from different sires and groups. This data was examined for quality control purposes but is not presented in this thesis.

Subsequently, accuracies were estimated for the entire (sub)population by calculating the correlation of the mean of each marginal density function with its corresponding true value from the simulation.

Table 5-3 Starting values

Initial values chosen for the two chains used to evaluate the burn-in period (chains 1 & 2) and for the subsequent analyses (chain 3). Description of the true population parameters are given in section 5.1.4.3.

	Chain 1	Chain 2	Chain 3
ψ	-4.6	2.6	0
ι	-4.6	2.6	0
a_ψ	-2.2	2.2	0
a_ι	-2.2	2.2	0
$\sigma^2_{e_\psi}$	0.6	1.4	2
$\sigma^2_{e_\iota}$	0.6	1.4	2
$\sigma_{e_{\psi\iota}}$	-0.4	0.4	0.3
$\sigma^2_{a_\psi}$	0.6	1.4	1.5
$\sigma^2_{a_\iota}$	0.6	1.4	1.5
$\sigma_{a_{\psi\iota}}$	-0.4	0.4	0.3
μ_ψ	-1.5	-1.1	-1
μ_ι	-1.5	-1.1	-1

5.1.4.3 Simulation studies

For the purpose of validation we assumed 'challenge' conditions for the simulation. Specifically we assumed that each epidemic is started by a single randomly chosen individual, called the index case, at t=0 in an otherwise naïve population and progresses through a series of independent infection events. The simulated populations consisted of N=2 000 paternal half-sib offspring from 100 sires for all scenarios. All parents were assumed to be unrelated. The simulation was run for a population with variation introduced in both underlying susceptibility and infectivity.

As stated above, breeding values and phenotypes on the underlying scale were assumed to be normally distributed. Specifically, the breeding values of the parental generation were sampled from N~(0, 1) for both underlying susceptibility and infectivity. The breeding values of the offspring generation were taken as the average of the parents plus a Mendelian sampling term taken from N~(0,0.5). The environmental deviations of the offspring were sampled from the normal distribution N~(0, 1), thus assuming a heritability h^2=0.5 for both underlying susceptibility and infectivity. The population mean for both underlying traits was set at -1.3. At the probability level this is approximately equal to 0.21 which is similar to the values used in all previous chapters. Susceptibility and infectivity were assumed to be independent for both breeding values and phenotypes.

The offspring for the standard population were distributed in groups of size n=10, 4 and 2 at random without reference to pedigree, giving 200, 500 and 1000 groups respectively. Each epidemic was run in closed groups, therefore no transmission was assumed between groups. The simulated epidemic followed a stochastic SI model, i.e. susceptible (S) individuals may become infected (I) and then remain infected. We assumed that infected individuals are immediately infectious. The transition of an individual from the S state to the I state occurs over a continuous time period but were observed only at discrete sampling times of length 0.5 units. Similarly to Chapter 2, on average, the rate of transition, from the state S to I, of a susceptible individual *j* due to an infected individual *k*, was assumed to take the following form,

$$\beta_{jk} = -ln(1 - H(\tau_j - t)g_j H(t - \tau_k s_k)f_k). \quad (5.8)$$

The other details of the simulated epidemic follows a stochastic Poisson process, as detailed in Chapter 2. The simulation was run up to time t=60 by which time prevalence was approximately 90%.

5.2 Results and Discussion

5.2.1 Evaluation of the program and its theoretical framework

The accuracies for susceptibility, infectivity and their breeding values, obtained when assuming that all the other parameters are known, are shown in Table 5-4. Given the stochastic nature of the simulations that generated the data, overall the accuracies displayed in Table 5-4 are reasonably high. The underlying infectivity had the lowest accuracy as could be expected due to the indirect nature in which it affects the observed data t_B and t_E. The marginal densities obtained for the population parameters are shown in Figure 5-1. The marginal densities for all the population parameters are distributed across the chance deviation of the sire and offspring samples from the distribution they were sampled from. It is therefore reasonable to assume, in addition to the relatively high accuracies for susceptibility, infectivity and their breeding values, that the algorithm is reliable and able to produce correct estimates for all the variates if sufficient information is provided.

One of the design variables implicit in the algorithm is the length of time between data collection points. This is important because it must be noted that one of the assumptions of the algorithm is that individuals infected at the same time do not affect each other. This assumption is met for example if the time period is sufficiently short such that only one individual becomes infected or there is a latency period which is longer than the observation interval. Therefore the disease of interest determines the appropriate length of observation period. Preliminary exploration of the simulated data suggested that the maximum observation period where this assumption holds true with the simulated data is approximately 5 units long. The observation used for the following analyses was 0.5 units. Preliminary results on the effect of observation interval size showed that increasing the interval size to a time period equal to 2 units had little effect on the estimates overall. However, a reduction in the accuracy of the estimates for susceptibility at the underlying level was observed for the last infected individuals. Thus better results are expected for slow spreading diseases.

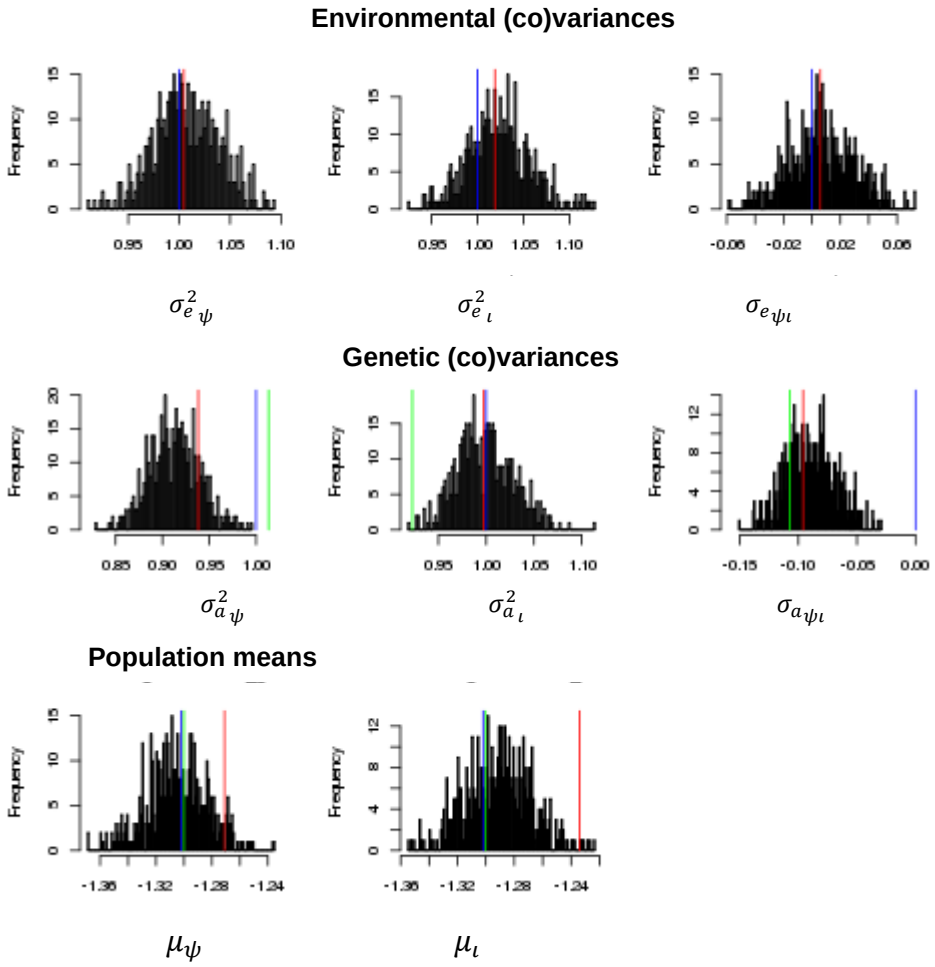

Figure 5-1 Marginal densities obtained when all other variables, except for the true infection times, are known

The red line indicates the offspring sample value, the green line the sire sample value and the blue line the distribution value they were sampled from in the simulation.

Table 5-4 Accuracy of breeding values a_ψ and a_ι and phenotypes ψ and ı when all other variables, except for the real infection time, are known

	a_ψ	a_ι	ψ	ı
Sires	0.88	0.89		
Offspring	0.73	0.74	0.75	0.64

5.2.2 Full sample

5.2.2.1 Burn-in and acceptance rate

The average acceptance rates are given in Table 5-5. As can be seen from the table, the acceptance rates for the genetic and environmental (co)variances are very low. It is possible to increase them by reducing the variance of their proposal distribution. However, attempting to do this resulted in poorer mixing and a reduction of the acceptance rates of the breeding values (results not shown).

The burn-in period is assumed to have been reached once both chains start to overlap and remain around the same area. Given that this state seems to be quickly reached, for most variables, it was deemed sufficiently cautiousto draw marginals from the last 10 000 of a chain of 100 000 samples. However, over the course of these studies it became clear that the chains for the environmental variance of infectivity and the population means were not stationary and remained dependent on the starting values even after two million cycles. Further work is required to resolve this issue. Given the extremely low acceptance rate for the variance components and the fact that the population means are dependent on the environmental (co)variance estimates further enquiries into different proposal distributions and/or prior assumptions may help to resolve this issue. Nonetheless, the work presented here provides valuable insights with regards to factors affecting the ability to estimate susceptibility and infectivity as comparisons are made within the same conditions.

Table 5-5 Acceptance rate per variable

	Acceptance rate %
τ	86
ψ	43
ι	71
a_ψ	16
a_ι	22
Environmental (co)variances	1
Genetic (co)variances	1
μ_ψ	20
μ_ι	37

5.2.2.2 Dependence of the parameter estimates on infection times

Table 5-6 shows the accuracy of susceptibility and infectivity estimates depending on infection time. These results confirm Bishop et al.'s (2012) hypothesis that time of infection would affect the accuracy of susceptibility estimates due to dynamic changes in infection pressure. Indeed, as can be seen in Table 5-6 the ability to estimate an individual's susceptibility increases with its infection time. Moreover, the ability to estimate infectivity is highest for the index cases and quickly drops off thereafter. Heuristically this makes a lot of sense as for example individuals that never became infected do not express infectivity. This is reflected in the estimates which are slightly negatively correlated with the true infectivity value assigned to these individuals in the simulation. It may therefore be worth enquiring how the estimates of such non-informative individuals affect the estimation of the population parameters and whether such censoring may be taken into account in the algorithm. It is also worth noting that although one might expect that the accuracy of the estimates of susceptibility and its breeding value to be the lowest for the index cases it is in fact lower for individuals that became infected early. This may be due to difficulties in disentangling the susceptibility of that individual from the infectivity of the index case. As more individuals become infected during the course of the

epidemic the sum of the infectivity of the infected individuals is likely to average out reducing the noise for estimating susceptibility. Similarly, at the start of the epidemic the number of available susceptible individuals would ensure that most index cases would start by infecting an individual with a susceptibility level at the higher end of the distribution, thus reducing the noise for estimating infectivity. These observations also confirm a hypothesis from a different perspective advanced in Chapter 1. There it was conjectured that when the epidemic follows an SI model the variance in susceptibility is scaled by I^2 and that in infectivity by $(1-I^2)$. It would therefore be easiest to capture these variance components at the end and at the start of the epidemic, respectively (see section 1.2.1.1.2., Figure 1-2).

Overall, the accuracy for infectivity, susceptibility and its breeding value are around 50% for informative individuals. The estimates of the infectivity breeding values of the sires, however, are down to 20%. This may be due to the relatively small number of informative individuals, i.e. index cases, per sire and/or to difficulties in estimating the variance components, as can be seen in Figure 5-2. Accuracies obtained when the variance components are known are shown in Table 5-7. These results indicate that the accuracy of the estimates for the breeding values of infectivity could be in line with the other accuracies if it is possible to estimate the variance components correctly. Work is currently being done to improve the variance component estimates. For example, a Metropolis within Gibbs algorithm is being developed assuming an Inverse Wishart rather than flat prior for the variance components. Moreover, enquiries are being made into excluding estimates from non-informative individuals with regards to either susceptibility or infectivity. Indeed, Ødegård et al. (Bangera *et al.* 2013; Ødegård *et al.* 2011) have successfully disentangled susceptibility from endurance with an MCMC algorithm of the Cure model which treats subgroups of individuals differently. It may also be possible that the burn-in period allowed was insufficient. In that case it may be possible to find a better set of proposal distributions with regards to the acceptance rates and to identify the appropriate burn-in period. Finally, if none of these steps help to improve the variance components estimates it might be necessary to estimate the variance components separately to be used with the MH algorithm to estimate the breeding

values. This type of two step approach is traditionally used in quantitative genetics. Indeed, breeding values are typically estimated using Best Linear Unbiased Prediction conditional on the variance components (Henderson 1975) after their estimation using Restricted Maximum Likelihood. A more detailed discussion regarding ongoing and future work to improve the algorithm is provided in Chapter 6.

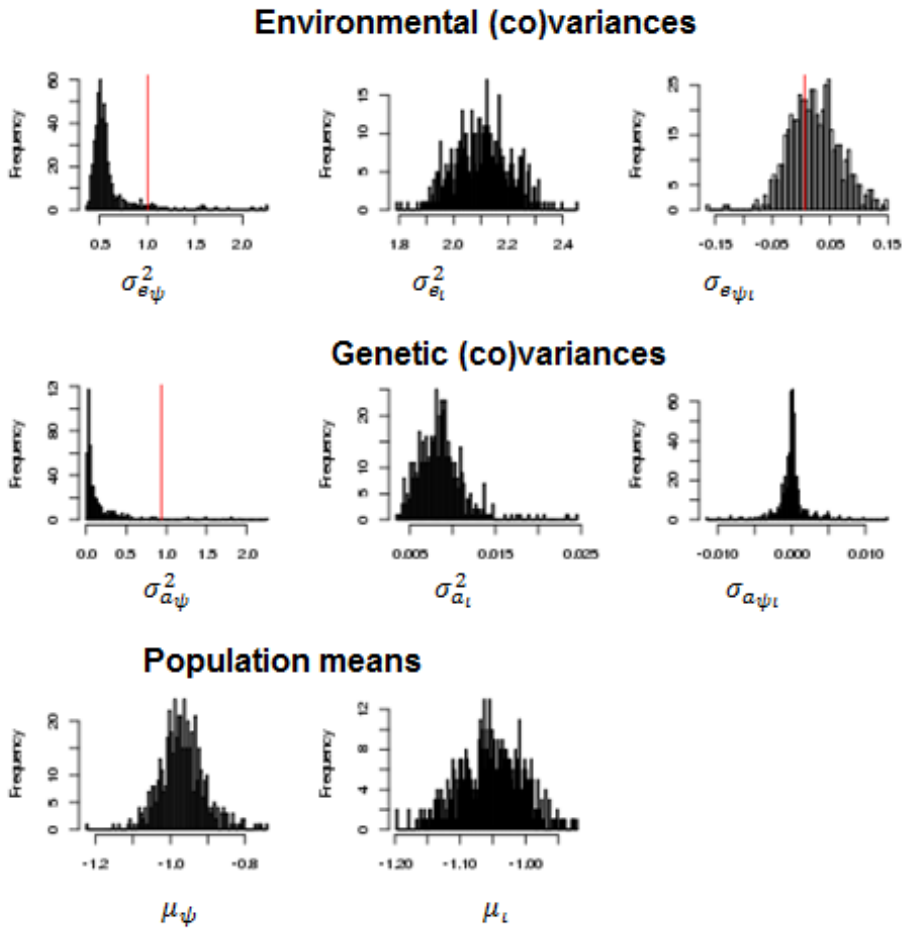

Figure 5-2 Marginal density functions for the population parameters

The true values are indicated by a red line where possible, where it is not indicated it fell outside the presented range. For the full set of true values and input values see Figure 5-1 and section 5.1.3.1, respectively.

Table 5-6 Accuracy of susceptibility and infectivity by infection time

	a_{ψ}	a_{ι}	ψ	ι
sires	0.59	0.20		
index	0.26	0.11	0.18	0.53
early	0.25	0.04	0.09	0.14
late	0.34	0.00	0.48	0.12
never	0.37	0.13	0.44	-0.09
overall	0.40	0.11	0.45	0.23

Table 5-7 Accuracy of susceptibility and infectivity when the variance components are known

	a_{ψ}	a_{ι}	ψ	ι
sires	0.59	0.45		
index	0.21	0.39	0.10	0.51
early	0.16	0.10	0.05	0.12
late	0.27	0.09	0.46	0.12
never	0.35	-0.03	0.42	-0.09
overall	0.34	0.19	0.45	0.21

5.2.2.3 Impact of group size on parameter estimates

Table 5-8 shows the accuracies for susceptibility and infectivity and their breeding values for populations of equal size N=2 000 split into groups of size two, four and ten. The results indicate a trade-off, between the ability to estimate susceptibility and the ability to estimate infectivity, depending on group size. Specifically, the ability to estimate susceptibility and corresponding breeding values seems to increase with group size. However, the ability to estimate infectivity is roughly equal for group sizes two and four and drops down for group size ten. The fact that susceptibility and corresponding breeding values are estimated with highest accuracy when the group size is large may be due to the fact that the averaging effect as well as the number of informative individuals, as described in the previous section, increases with group

size. For example the number of individuals which are not index cases increases from 50 to 90% from group size 2 to 10. Therefore, in order to estimate infectivity, on the other hand, there is trade-off as the averaging effect increases and the number of informative individuals decreases with group size. A compromise regarding group size is therefore indicated and all further analyses were performed on populations divided in groups of size four. The only pattern observed regarding the population parameters was an increase in the environmental variance of the underlying infectivity as group size increased (results not shown).

Table 5-8 Accuracy of susceptibility and infectivity estimates by group size

Index cases are assumed to be the most informative individuals for the estimates of infectivity and its breeding values and individuals which never became infected are assumed to be most informative for the estimates of susceptibility and its breeding values.

*Late infected individuals were more informative in this scenario with an accuracy of 0.48.

	Group size	Accuracy		
		Overall	Most informative	Sires
a_ψ	2	0.14	0.22	0.27
	4	0.40	0.37	0.59
	10	0.43	0.60	0.58
a_ι	2	0.11	0.17	0.20
	4	0.11	0.11	0.20
	10	0.08	0.14	0.16
ψ	2	0.23	0.11	
	4	0.45	0.44*	
	10	0.52	0.72	
ι	2	0.23	0.45	
	4	0.23	0.53	
	10	0.10	0.38	

5.3 Conclusions

An MCMC algorithm was developed to infer breeding values in susceptibility and infectivity from longitudinal binary infectious disease data. Through the use of simulated data it was demonstrated that the algorithm is capable of estimating all parameters correctly if sufficient information is provided. However, there are convergence issues for the population parameter estimates when all parameters are being sampled. Nonetheless, it is possible to infer values for susceptibility and its corresponding breeding values and for infectivity with reasonable accuracy using the algorithm as it stands. It is also possible to infer breeding values for infectivity to a similar level of accuracy if the variance components are known. Following implementation of the algorithm, results suggested that the accuracy of parameter estimates is dependent on infection time as hypothesized by Bishop et al. (2012). In particular, accuracy seems to increase with infection time for susceptibility estimates and corresponding breeding values whilst the accuracy for infectivity and corresponding breeding values is highest for the individuals which start the epidemic. There also seems to be a trade-off in terms of group size as the accuracy of susceptibility and corresponding breeding values increases with group size whilst the accuracy of infectivity seems to be lower for a larger group size. Future work will focus on the ability to estimate variance components accurately.

5.4 References

Bangera, R., J. Ødegård, H. M. Nielsen, H. M. Gjoen and A. Mortensen, 2013 Genetic analysis of vibriosis and viral nervous necrosis resistance in Atlantic cod (Gadus morhua L.) using a cure model. Journal of Animal Science **91:** 3574-3582.

Bishop, S. C., A. B. Doeschl-Wilson and J. A. Woolliams, 2012 Uses and implications of field disease data for livestock genomic and genetics studies. Front Genet **3:** 114.

Gitterle, T., J. Ødegård, B. Gjerde, M. Rye and R. Salte, 2006 Genetic parameters and accuracy of selection for resistance to White Spot Syndrome Virus (WSSV) in Penaeus (Litopenaeus) vannamei using different statistical models. Aquaculture **251:** 210-218.

Henderson, C. R., 1975 BEST LINEAR UNBIASED ESTIMATION AND PREDICTION UNDER A SELECTION MODEL. Biometrics **31:** 423-447.

Muir, W. M., 2005 Incorporation of competitive effects in forest tree or animal breeding programs. Genetics **170:** 1247-1259.

Ødegård, J., T. Gitterle, P. Madsen, T. H. E. Meuwissen, M. H. Yazdi *et al.*, 2011 Quantitative genetics of taura syndrome resistance in pacific white shrimp (penaeus vannamei): a cure model approach. Genetics Selection Evolution **43**.

Ødegård, J., I. Olesen, B. Gjerde and G. Klemetsdal, 2006 Evaluation of statistical models for genetic analysis of challenge test data on furunculosis resistance in Atlantic salmon (Salmo salar): Prediction of field survival. Aquaculture **259:** 116-123.

Ødegård, J., I. Olesen, B. Gjerde and G. Klemetsdal, 2007 Evaluation of statistical models for genetic analysis of challenge-test data on ISA resistance in Atlantic salmon (Salmo salar): Prediction of progeny survival. Aquaculture **266:** 70-76.

Chapter 6. General discussion

6.1 Contributions of the thesis

The aim of this thesis was to develop improved methods to estimate the host genetic contribution to the epidemiology of infectious diseases. Ultimately better estimation methods are paramount for implementing genetic selection as part of control strategies for infectious diseases in livestock. The starting hypothesis of the thesis was that current tools may be missing important heritable variation as little is known regarding the genetic contribution of host infectivity and it is currently ignored. This stands in contrast with abundant evidence that heterogeneity in host infectivity, super-shedders being an extreme example, is ubiquitous and can profoundly impact upon disease prevalence in the population (Lloyd-Smith *et al.* 2006). In this way, Chapter 2 identified that failing to include the heritable effect of other individuals on the disease status of individuals when analysing data from field studies does indeed provide no information regarding host infectivity and may thus result in substantial genetic variation being missed. For example, had the QTL, explaining 98% of the additive genetic variation in susceptibility to pancreatic necrosis in Salmon, found by Houston et al. (2010) affected infectivity rather than susceptibility it would probably have been overlooked. Chapter 2 demonstrates that it is however possible to capture some of the genetic variation in host infectivity by using an Indirect Genetic Effects (IGE) model. Moreover, this additional genetic variance does not come at the expense of obtaining reliable estimates for susceptibility.

Although the IGE model had some utility with regards to estimating breeding values in infectivity, it was far from perfect with only a fraction of the available genetic variation in infectivity being captured. Nonetheless, in Chapter 2, it was determined that the basic reproduction number R_0, and thus the risk and severity of outbreaks, is reduced when selecting for lower infectivity estimated breeding values using the IGE model. This would be particularly relevant in instances where there is substantially more variation in infectivity compared to susceptibility and selection would have otherwise not been feasible. It is as yet unknown whether infectivity harbours

substantial genetic variation but this work provides the first tools to address these questions.

Chapter 3 demonstrates that the bias, accuracy and impact of selection of the IGE model used for genetic analysis binary disease data could be slightly improved by specifying the individuals contributing to the indirect effect with an incidence matrix. These results further strengthened the results obtained in Chapter 2. However, the results of this study also pointed out severe shortcomings in trying to take account of disease dynamics within conventional quantitative genetics mixed model framework and software (ASReml, Gilmour *et al.* 2006).

One of the main issues encountered was that the relationship between the observed binary data and the epidemiological parameters of interest was unknown. A genetic-epidemiological link function was therefore derived in Chapter 4, from quantitative genetics and epidemiological theory, which explicitly links susceptibility and infectivity to the observed binary data. Using this function it was demonstrated that ignoring either exposure dynamics or genetic heterogeneity in susceptibility and infectivity would provide a poor link to binary disease data. The derived function provided insights as to which link functions are appropriate or generate bias under what conditions. However, the link function cannot be integrated into existing software. A Markov Chain Monte Carlo (MCMC) algorithm was therefore developed, in Chapter 5, to estimate breeding values in susceptibility and infectivity from repeated binary disease data.

The algorithm developed in Chapter 5 is subject to ongoing work as there are current difficulties to estimate the variance components. Nonetheless, it is possible to obtain reasonably accurate estimates of susceptibility and infectivity at the phenotypic level and of the breeding values of susceptibility using the algorithm as it stands. If the variance components are known then the accuracy for the breeding values of infectivity increases to a similar magnitude as that for other parameters. Proposed methods to improve the estimation of the variance components are fully discussed in the next section.

In conclusion, the thesis advances and develops a novel approach to the analysis of binary infectious disease data, which makes it possible to capture genetic variation in both host susceptibility and infectivity. This approach has been refined in this thesis and is subject to ongoing work to make those estimates increasingly accurate. These breeding values will provide novel opportunities for genome wide association studies and may lead to novel genetic disease control strategies tackling not only host resistance but also the ability to transmit infectious agents.

6.2 Further improvement of the MCMC algorithm

One of the key outcomes of this thesis is the development of an MCMC algorithm for estimating genetic parameters associated with host susceptibility and infectivity. The main area for future improvement of this algorithm identified in Chapter 5 is estimation of the variance components. It is possible, given the extremely low acceptance rate for the variance components that the burn-in period allowed was insufficient. However, the space being explored by each chain remained the same after 2 million cycles even though differences in this space were observed between chains i.e. they did not converge. Moreover, simply identifying a more suitable set of proposal distributions in terms of acceptance rates is no simple task as changes in the proposal distribution of one parameter seemed to change the acceptance rates of all other parameters in complex and unpredictable ways. However, the proposal distribution for the variance components was log-normal with the previous estimate taken as the mean on the underlying level. Therefore the closer an estimate comes to zero the less likely it is to move away from it. Furthermore, if a variance component estimate is equal to zero then the requirement for a positive definite matrix cannot be met thus limiting the number of samples which may be accepted. The chain therefore remains ‘stuck’ around zero for that component and around an arbitrary value for the other components. Constraining a uniform prior for the variance components between $]0,A]$ with $A \rightarrow \infty$ leads to the opposite problem of inflated variance estimates (Gelman 2006). Preliminary results taking such an approach with the algorithm developed in Chapter 5 led to very inflated environmental (co)variance

estimates and genetic (co)variance estimates close to zero (results not shown). Due to such difficulties using uninformative priors for variance components as well as for reasons of computational ease, inverse Wishart priors are often assumed for the variance components. A Metropolis-within-Gibbs algorithm using this assumption is currently being implemented. This should not only help to resolve the problems with estimating variance components but should also render the program faster and more efficient.

One of the main findings of Chapter 5 is that the accuracy of the estimated values and breeding values for susceptibility and infectivity depends on time of infection as hypothesized by Bishop et al. (2012). However, estimates for all individuals contribute equally to the estimation of the population parameters regardless of the fact that some estimates are based on very little information. For example, individuals which never became infected never express infectivity, attempting to estimate it results in estimates which are slightly negatively correlated with the true infectivity values from the simulation. The algorithm is currently being adapted to take such censored information regarding susceptibility or infectivity into account. In principle sampling these censored values can be done appropriately but explicitly dealing with the censored data will be more efficient. Finally, the accuracy of all estimates may be improved by optimising the experimental design as discussed in the following section.

6.3 Implementation of findings to real data

Following the assumptions and the findings of this thesis, in order to estimate breeding values in host susceptibility and infectivity using the MCMC algorithm developed in Chapter 5, the following criteria are required.

The disease of interest should be relatively slow to spread, i.e. low R_0, providing sufficient time to make observations before the entire group is infected. According to Woolliams (2012) endemic diseases with a low R_0 are probably the most effective

targets for control strategies including genetic selection. Indeed, a small reduction in risk and severity of outbreaks would have very little impact on the management of a virulent disease but it would have a big impact if that small change brings the disease below the threshold for major epidemics.

Ideally diagnostic tests should have a high sensitivity and specificity as it was demonstrated by Bishop and Woolliams (2010) that imperfect test sensitivity and specificity cause underestimation of heritability. Moreover, they need to be relatively fast to implement and not too intrusive so that the test may be repeated frequently and preferably be inexpensive. Relatively non-intrusive tests such as faecal egg counts, cloacal/nasal swabs, skin tests and ELISA are available for a wide range of diseases. The cost in terms of time and economics as well as the performance of the diagnostic test depend on the disease and species of interest and would have to be weighed against the relative benefits to be gained.

In order to be able to estimate breeding values in infectivity, natural transmission of the disease has to be able to occur and data must originate from a large number of contact groups. Thus for example data of an endemic disease from a large number of farms or a large number of separately housed groups within a farm would be suitable. However, at the moment the algorithm is suited to diseases following a relatively simple SI model in closed groups. It would therefore only be applicable to a limited number of diseases and livestock species. Further work is being carried out at The Roslin Institute to expand the algorithm for a greater range of epidemiological models and contact networks.

Prior to implementing the MCMC algorithm with real data, some consideration should be given to optimal design strategies. Questions of optimal design were investigated by Bijma (2010) with regards to Indirect Genetic Effects (IGE) models. Although the MCMC algorithm developed in Chapter 5 does not include an IGE model in its standard linear form, the cumulative force of infection could be considered as an IGE model with an additional non-linear term (see Chapter 4). In

this way, some of Bijma's (2010) findings may be an appropriate starting point for further investigations.

The results of (Bijma 2010) are presented depending on the dilution of the IGE, where full dilution is defined as a situation where the IGE of an individual reduces with group size and no dilution as a situation where the IGE is independent of group size. Throughout this thesis contact rate is assumed to be equal to 1 and the transmission rate is not density dependent. In this way one might intuitively assume no dilution. However, the relative contribution of one individual to the infection status of others decreases with group size as the probability of being the individual to transmit the infection is decreased as more individuals become infected. Thus dilution increases over the course of the epidemic. This property is reflected in the results obtained throughout the thesis. Hence, for the purpose of this discussion, strong dilution will be assumed.

Bijma (2010) explored the effect of group composition with regards to relatedness by comparing two ends of the scale. At one end of this scale, individuals were grouped at random with regards to family, as assumed in this thesis. On the other end each group was composed of two families. The two families per group design seemed to attenuate the requirement for large family sizes for the estimation of IGEs. However, using such a design with the MCMC algorithm, with a group size of 4, seemed to provide similar or worse estimates in general (results not shown). This is in line with the findings of Bijma (2010) that there would be no advantage in using the two families per group design in the event of strong dilution when the number of individuals rather than groups is the limiting factor. However, the estimates for susceptibility and its BV for individuals infected early provided an exception with their accuracies (25 and 34% respectively) increased above that of the index cases (results not shown). In other words, the two families per group design, whilst it did not confer any advantage in general, did help to disentangle the susceptibility of the first individual to be infected from the infectivity of the index case. Furthermore, one of the advantages of limiting the number of families per group and therefore increasing the number of relatives in the group is that it reduces the effective size of

the group. However, this limits variation within the group and therefore limits the averaging out of the indirect effect as group size increases. Group composition is therefore expected to be of greater importance as group size increases, as demonstrated by Ødegård and Olesen (2011).

In the case of strong dilution, Bijma's (2010) findings suggest that the optimum group size would be as small as possible for the indirect effect, i.e. infectivity, whereas the optimum group size for the direct effect, i.e. susceptibility, would be much larger. This is in line with the results of Chapter 5. The only scenario in which the indirect effect benefitted from larger groups in the findings of Bijma (2010) was in the case that the number of groups was the limiting factor and increasing group size corresponds to an increase in total population size. Thus, as might be expected, the optimum design would entail as large a population as possible. This may only be achievable to obtain with field rather than experimental data, as previously mentioned. A potential solution might be to have a smaller experimental setting where relatives of individuals in the field are placed in small groups. In this way, each family should have information from small groups, allowing the estimation of infectivity, as well as larger groups, allowing the estimation of susceptibility.

Given equally sized groups, requirements for a large population size, large family size and small groups suggest that the implementation of the algorithm as it stands may be feasible with e.g. poultry, pig and/or fish data. The large fecundity of male and female fish makes fish species ideal candidates. However, to avoid the stress of removing fish from their environment, challenge studies on fish species traditionally record their mortality (Ødegård *et al.* 2011) whereas time of infection would be the trait of interest in order to estimate transmission parameters. Nonetheless, diagnoses exist for some diseases affecting fish species which are non-intrusive and repeatable such as observing skin pigmentation for *Philasterides dicentrarchi.* Pig and poultry are equally good candidates as large paternal half-sib family sizes are available and it is not uncommon for them to be housed in small closed groups. For example chicken breeding lines may be housed in cages of four.

6.4 Future opportunities

The development of the MCMC algorithm from Chapter 5 has generated several new projects at The Roslin Institute. Proposals have been submitted and pilot projects initiated to apply the algorithm to *Philasterides dicentrarchi* in turbot and Coccidiosis and Marek's disease in poultry. This would entail adapting the algorithm for more complex epidemiological models and population structure. As there may be variation in recovery rate and/or latency it may be necessary to consider infectiousness or a decomposition of its components, duration of infectious period, infectivity and contact rate, to be the parameters of interest. Decomposing the components of infectiousness may provide interesting insights regarding the causes of heterogeneity in infectiousness. Moreover, in the future work may need to be done to expand the model to allow for heterogeneous contact rates/ networks.

It is not yet known to which extent infectivity is heritable as to our knowledge the work developed in this thesis provides the first tools to address this question. Moreover, as mentioned above the algorithm still requires further development before being applicable to a range of diseases. Nonetheless, it is possible to capture some of the variation in infectivity, if present, with the IGE model presented in Chapter 3. Moreover, several promising options are being explored to overcome the convergence problems of the MCMC algorithm. It is therefore plausible that if the major QTL found by Houston et al. (2010) had affected infectivity rather than susceptibility it would now be possible to use methods described in this thesis to detect it and provide EBVs.

Overall, the work developed in this thesis provides a starting point for a promising new area of enquiry. Indeed, providing that infectivity is sufficiently heritable for the above mentioned studies and that the algorithm may be adequately expanded to provide reasonably accurate breeding values these could then be used as part of genome wide association studies. Identifying genomic markers for infectiousness and/or any of its components could lead to novel genetic disease control strategies

tackling not only host resistance but also eliminating super-spreaders. Furthermore, selective breeding is only one of the possible applications and other applications may be relevant to humans too. For example, breeding values may be used to target super-spreaders for vaccination and/or treatment. Indeed, Matthews et al. (2006) demonstrated that the R_0 of E. coli in cattle could be reduced to below 1 by preventing infection in the individuals with the highest 5% of infectiousness as measured by bacterial counts. Moreover, using these markers it may be possible to look for causal genes and a deeper understanding of the mechanisms underpinning infectiousness. This may lead for example to the development of chemical treatment targeting disease transmission as a novel way to combat endemic diseases.

6.5 References

Bijma, P., 2010 Estimating Indirect Genetic Effects: Precision of Estimates and Optimum Designs. Genetics **186:** 1013-1028.

Bishop, S. C., A. B. Doeschl-Wilson and J. A. Woolliams, 2012 Uses and implications of field disease data for livestock genomic and genetics studies. Frontiers in genetics **3:** 114.

Bishop, S. C., and J. A. Woolliams, 2010 On the Genetic Interpretation of Disease Data. PLoS ONE **5:** e8940.

Gelman, A., 2006 Prior distributions for variance parameters in hierarchical models(Comment on an Article by Browne and Draper). Bayesian Analysis **1:** 515-533.

Gilmour, A. R., B. J. Gogel, B. R. Cullis and R. Thompson (Editors), 2006 *ASReml User Guide Release 2.0*. VSN International Ltd, Hemel Hempstead, UK.

Houston, R. D., C. S. Haley, A. Hamilton, D. R. Guy, J. C. Mota-Velasco *et al.*, 2010 The susceptibility of Atlantic salmon fry to freshwater infectious pancreatic necrosis is largely explained by a major QTL. Heredity **105:** 318-327.

Lloyd-Smith, J. O., S. J. Schreiber and W. M. Getz, 2006 Moving beyond averages: Individual-level variation in disease transmission, pp. 235-258 in *Mathematical Studies on Human Disease Dynamics: Emerging Paradigms and Challenges*, edited by A. B. GUMEL. American Mathematical Society, Providence.

Matthews, L., J. C. Low, D. L. Gally, M. C. Pearce, D. J. Mellor *et al.*, 2006 Heterogeneous shedding of Escherichia coli O157 in cattle and its implications for control. Proceedings of the National Academy of Sciences of the United States of America **103:** 547-552.

Ødegård, J., M. Baranski, B. Gjerde and T. Gjedrem, 2011 Methodology for genetic evaluation of disease resistance in aquaculture species: challenges and future prospects. Aquaculture Research **42:** 103-114.

Ødegård, J., and I. Olesen, 2011 Comparison of testing designs for genetic evaluation of social effects in aquaculture species. Aquaculture **317:** 74-78.

Woolliams, J., 2012 Influence of genetics and inbreeding on disease. In Practice **34:** 196-+.

Appendices

Appendix 1. Derivation of transmission parameter from first principles

We define the probability of a susceptible individual *j* to become infected upon contact with an infected individual k as the product of the susceptibility of *j* (g_j) with the infectivity of *k* (f_k), with the assumption that susceptibility and infectivity are independent. Let c_{jk} be the expected number of contacts between individuals *j* and *k* per time unit. The probability of a susceptible individual *j* to avoid getting infected per time unit will therefore be equal to

$$\prod_{k=1}^{n-1}(1-g_j X_{f,k} f_k)^{c_{jk}}. \tag{A.1}$$

The indicator trait $X_{f,k}$ is equal to one if *k* is infected and zero otherwise.

The probability *dq* of a susceptible individual *j* to become infected during a sufficiently short time period *dt* such that $X_{f,k}$ does not change for all individuals *k*, is therefore:

$$\begin{aligned} dq &= 1-\left(\prod_{k=1}^{n-1}(1-g_j X_{f,k} f_k)^{c_{jk}}\right)^{dt} \\ &= 1-\exp\left(dt\sum_{k=1}^{n-1} c_{jk}\ln(1-g_j X_{f,k} f_k)\right). \end{aligned} \tag{A.2}$$

Using the property, $e^x = 1+x+x^2/2!+x^3/3!+\ldots$, dividing by *dt* and taking *lim* $dt \rightarrow 0$, we obtain the rate of infection for one individual or force of infection,

$$\frac{dq}{dt} = -\sum_{k=1}^{n-1} c_{jk} ln(1-g_j X_{f,k} f_k). \tag{A.3}$$

Hence, the change in the number of susceptible individuals over a time period dt is given as:

$$\frac{dS}{dt} = \sum_{j=1}^{n} \sum_{k=1}^{n-1} c_{jk} ln(1 - X_{g,j} g_j X_{f,k} f_k). \quad \text{(A.4)}$$

The indicator trait $X_{g,j}$ is equal to one if j is susceptible and zero otherwise.

The pairwise transmission parameter β_{jk} is defined as the rate at which a susceptible individual j will become infected upon contact with an infected individual k. In this way,

$$\beta_{jk} = -ln(1 - X_{g,j} g_j X_{f,k} f_k). \quad \text{(A.5)}$$

Note that for small values of g and f this may be approximated by $\beta_{jk} = X_{g,j} g_j X_{f,k} f_k$.

Appendix 2. Derivation of variance in disease presence

Assuming that disease presence is distributed according to equation (2.4) and that the environmental component is independent from all other components, the variance in disease presence can be expressed as follows:

$$\begin{aligned}\sigma^2 &= \mathrm{cov}(b_1 g + b_2 \sum_{m=1}^{p} f_m + e, b_1 g + b_2 \sum_{m=1}^{p} f_m + e) \\ &= b_1^2 \sigma_g^2 + 2 b_1 b_2 \mathrm{cov}(g, \sum_{m=1}^{p} f_m) + b_2^2 \mathrm{var}(\sum_{m=1}^{p} f_m) + \sigma_e^2 \end{aligned} \quad \text{(A.6)}$$

Assuming that the number of individuals which have been infected p is a random variable and given independence between input susceptibility and infectivity,

$$\begin{aligned}\mathrm{cov}(g, \sum_{m=1}^{p} f_m) &= \mathrm{E}_p\left(\mathrm{cov}(g, \sum_{m=1}^{p} f_m \mid p)\right) + \mathrm{cov}_p\left(\mathrm{E}(g), \mathrm{E}(\sum_{m=1}^{p} f_m \mid p)\right) \\ &= \mathrm{E}(p)\mathrm{cov}(g, f) + 0 \\ &= 0 \end{aligned} \quad \text{(A.7)}$$

$$\begin{aligned}\mathrm{var}(\sum_{m=1}^{p} f_m) &= \mathrm{E}_p\left(\mathrm{var}(\sum_{m=1}^{p} f_m \mid p)\right) + \mathrm{var}_p\left(\mathrm{E}(\sum_{m=1}^{p} f_m \mid p)\right) \\ &= \mathrm{E}(p)\mathrm{var}(f) + \mathrm{E}^2(f)\mathrm{var}(p) \\ &= \bar{p}\sigma_f^2 + \bar{f}^2 \sigma_p^2. \end{aligned} \quad \text{(A.8)}$$

Incorporating equations (A.7) and (A.8) into equation (A.6) we obtain equation (2.5).

Appendix 3. Impact of model parameters on prevalence profiles

The desired outcome of control of infectious diseases through selection is a reduction in prevalence. Moreover, genetic parameter estimates of disease traits depend on disease prevalence (Bishop and Woolliams 2010). The impact of the genetic model parameters on prevalence profile was therefore examined.

Impact of mean susceptibility/infectivity on prevalence profiles

In order to examine the impact of different levels of infectivity or susceptibility on disease prevalence in our model we first ran simulations for homogeneous populations with two levels of infectivity f and susceptibility g. Specifically, for both infectivity and susceptibility the high level equals 0.4 and the low level 0.04. From Figure S1 it is clear that populations with different degrees of susceptibility/infectivity have different prevalence profiles. Note that, populations with a high level of susceptibility and low infectivity had the same expected prevalence over time as populations with low susceptibility and high infectivity (cf. Figure S1). In other words, different levels of infectivity or susceptibility have the same impact on disease prevalence in this model.

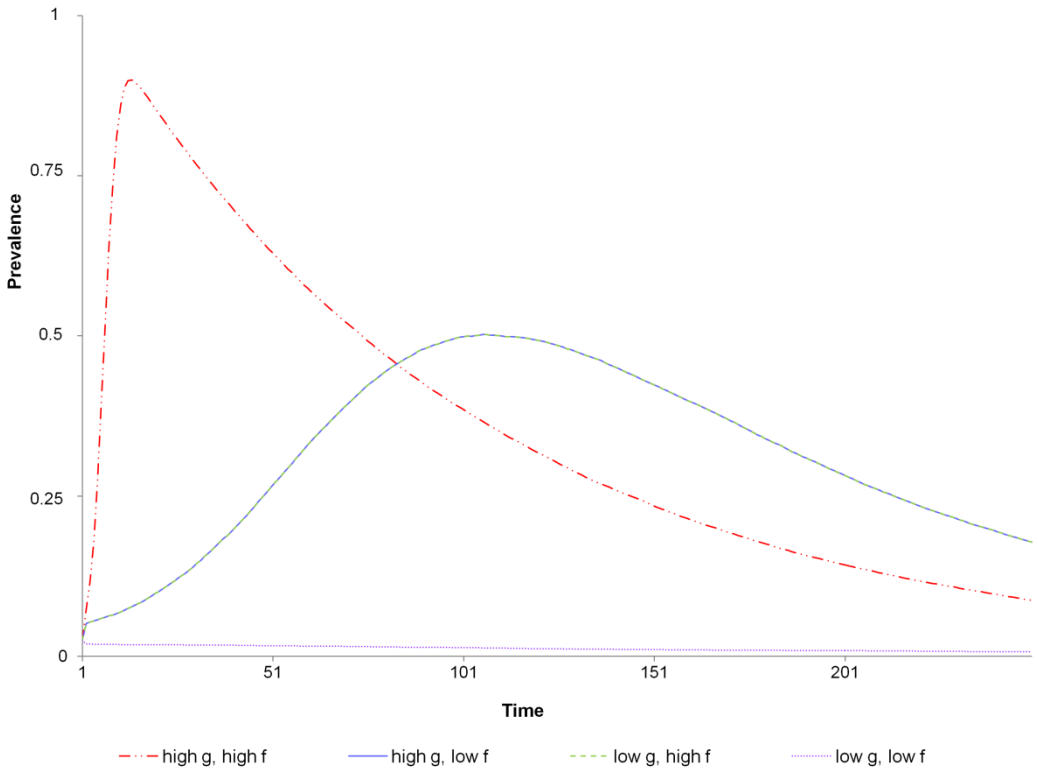

Figure S 1 Predicted disease prevalence over time

Homogeneous population for suceptibility (high g = 0.4, low g = 0.04) and infectivity (high f = 0.4, low f = 0.04). Population consists of 500 groups of size 40 as in Table S1. Prevalence was averaged over all groups over three iterations. Probability of disease emerging in a group was 0.38 in the population with low susceptibility and infectivity and 1 for the other populations. The expected course of the epidemic is identical for high infectivity/low susceptibility and low infectivity/high susceptibility.

Table S 1 Population structure parameters

Group size n	10	40	400
# Sires *s*	125	500	5000
Family size	40	40	40
Population size *N*	5 000	20 000	200 000
# Groups	500	500	500

Impact of variation in susceptibility and/or infectivity

The impact of variation in susceptibility and/or infectivity on disease prevalence at different stages of the epidemic for the different genetic models (i.e. bi-allelic vs multiple alleles; symmetric vs skewed) and different group sizes was examined. Figures S2 & S3 show the prevalence profile for populations, consisting of different group sizes, with variation introduced in either susceptibility or infectivity, neither or both traits using the skewed multiple allele and symmetric bi-allelic models respectively. Underlying genetic architecture and frequency distribution, however, had little impact on the time course of the epidemic (cf. Figure S2 & S3). Group size had the highest impact as with increasing group size the epidemic progressed faster towards its maximum prevalence and this maximum prevalence was increased (cf. Figure S2 & S3). The introduction of variation in susceptibility/infectivity had little impact on prevalence profiles although it slightly decreased disease peak prevalence. For all group sizes, the impact of heterogeneity was strongest when there was variation in both susceptibility and infectivity.

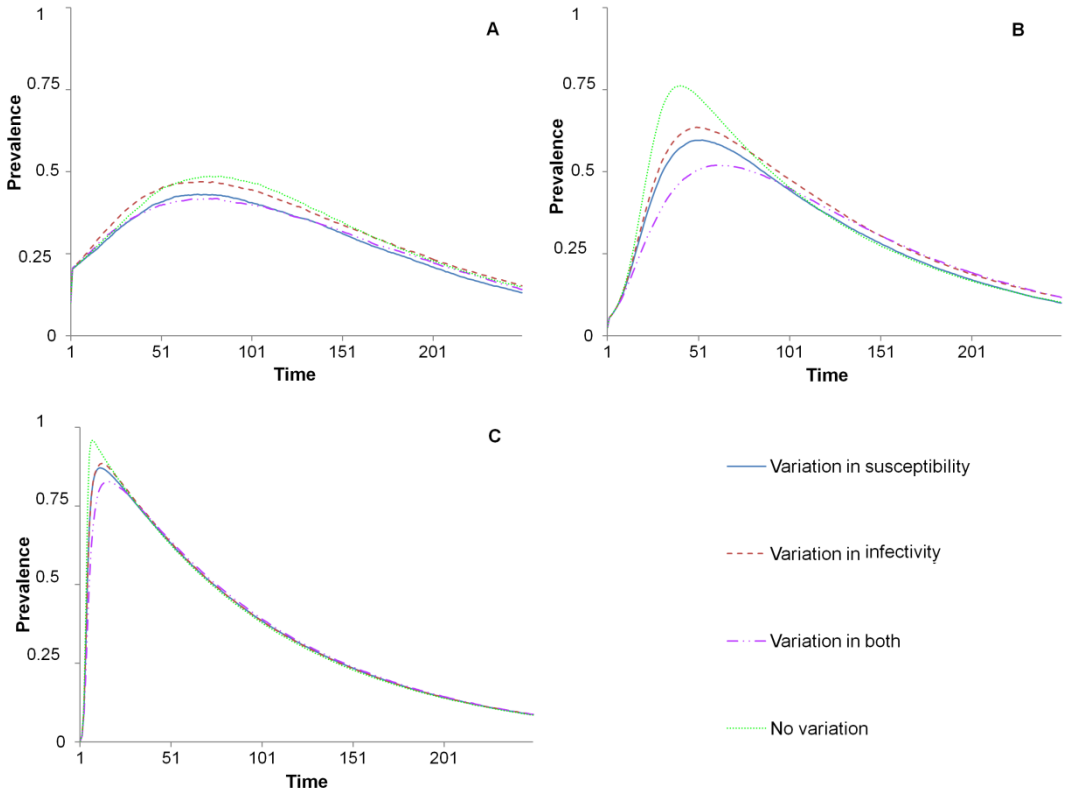

Figure S 2 Disease prevalence over time assuming many underlying alleles of varying effect coding for susceptibility or infectivity and a skewed distribution

Parameters as in Table 1-2. Population structure parameters as in Table S1. A) group size of 10 B) group size of 40 C) group size of 400.

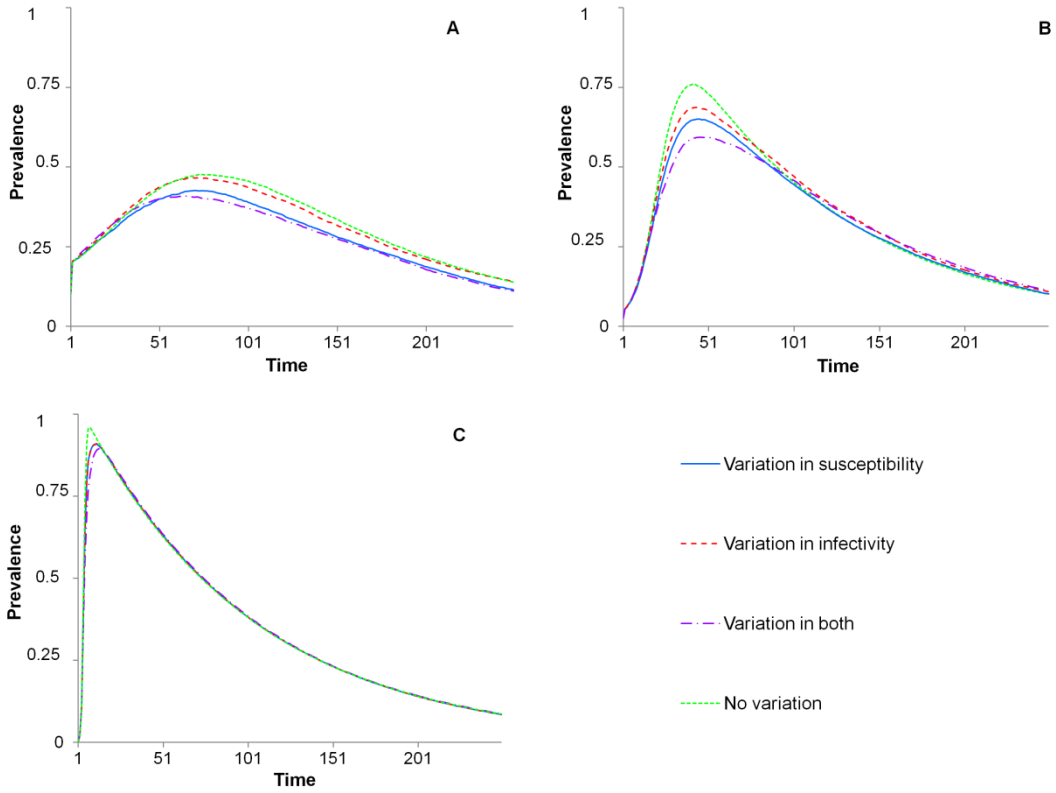

Figure S 3 Disease prevalence over time assuming two alleles code for susceptibility or infectivity and a symmetrical distribution

Parameters as in Table 1-2. Population structure parameters as in Table S1. A) group size of 10 B) group size of 40 C) group size of 400.

Appendix 4. Impact of a logistic regression on variance estimates and selection response

Generalized linear mixed models (GLMM), linking a linear mixed model with a non-linear link function, such as a logistic regression, are often used for the genetic analysis of binary data . It is therefore of interest to assess whether the use of a non-linear link function would alter the main messages of this paper.

For this purpose, the data from the population with variation in susceptibility and infectivity following a skewed multiple allele distribution was analysed with the conventional and the indirect genetic effect (IGE) model (equations 2.2 and 2.3) using a logistic link function. The variance estimates obtained from these analyses are displayed in Table S2. Similarly to the linear model without a link function, the animals with the lowest ten percent of estimated breeding values (EBVs) obtained from these analyses were then selected. The mean true values of infectivity, susceptibility and basic reproduction number R_0 for each selected subpopulation are displayed in Table S3.

Table S 2 Variance estimates using a logistic link function

Estimates averaged over ten replicates. Parameters as in Table 2, 10000 groups of size 10. Values ± standard error.

Model	**Conventional:**	**IGE**:		
		Direct	Indirect	Direct-Indirect
Component	σ_A^2	σ_D^2	σ_S^2	σ_{DS}
Estimate	0.178±0.004	0.210±0.005	0.009±0.001	0.024±0.002

Similarly to the results obtained without a link function, analysis with the IGE model with a logistic link function obtains a variance estimate for the direct effect which is approximately of the same magnitude as that obtained with the conventional model as well as a smaller yet significant variance for the indirect effect (see Table S2).

Again, similarly to the results obtained without a link function, selecting on conventional or direct EBVs results in a reduction in mean susceptibility only whereas selecting on indirect EBVs reduces both mean susceptibility and infectivity (see Table S3). The greatest reduction in R_0 is also obtained by selecting with an index of direct and indirect EBVs (see Table S3).

Table S 3 Mean susceptibility and infectivity following selection using the conventional animal model or the Indirect Genetic Effects model with a logistic link function

Population with variation in both infectivity and susceptibility following a skewed multiple allele genetic architecture. 10000 groups of size 10. Proportion selected was 0.10. Values ± standard error when greater than 0.005.

Selection			Mean susceptibility	Mean infectivity	R_0
None			0.22	0.22	4.46
Conventional animal effect		EBV	0.10	0.22	2.14±0.04
Direct effect		EBV_D	0.11	0.21	2.06±0.03
Indirect effect		EBV_s	0.13	0.18	2.13±0.06
Index	$Ix=EBV_D+\bar{p}$ (n-1)	EBV_s	0.11	0.20	2.03±0.03

It must be pointed out that the variance estimates in Table S2 are clearly not on the same scale as for the parameters underlying the data hence a logistic link function is not an appropriate transformation (see Table 2-2). Moreover, the direct-indirect covariance estimate is much larger than the indirect variance estimate and selection on the indirect EBVs obtained with a logistic link function results in an even greater reduction in mean susceptibility and less reduction in mean infectivity than its linear equivalent (see Tables S2 and 1-7). This would indicate that the use of a logistic link function perhaps aggravates the bias created by the interaction between expression of infectivity and susceptibility. Besides, selection on all except the indirect EBVs results in a greater reduction in R_0 when analysing without a logistic link function.

Bishop, S. C., and J. A. Woolliams, 2010 On the Genetic Interpretation of Disease Data. Plos One **5**.